Master Thesis

Seasonality in cholera dynamics in North-East India and Bangladesh

A rainfall-driven model explains the wide range of patterns of an infectious disease in endemic areas

Author:
Theo Baracchini
EPFL

Supervisors:
Professor **Mercedes Pascual**
Department of Ecology and Evolutionary Biology - University of Michigan, USA

Professor **Aaron A. King**
Department of Ecology and Evolutionary Biology - University of Michigan, USA

Senior scientist **Enrico Bertuzzo**
Laboratory of Ecohydrology - EPFL, Switzerland

Thesis director:
Professor **Andrea Rinaldo**
Laboratory of Ecohydrology - EPFL, Switzerland

University of Michigan - USA
-
Spring 2014

Acknowledgements

The success of a project is rarely the work of a person alone but depends on the encounters and guidance of many others. This master thesis is no exception, therefore I would like to take this opportunity to thank those who made it possible.

First and foremost, I would like to express my appreciation and thanks to my two supervisors, Prof. Mercedes Pascual and Prof. Aaron A. King, for their assistance and dedicated involvement in every step of this study. It has been very enriching and exciting to study and learn from such sharp and complementary minds. I am grateful to have had this privilege. Thank you very much for your patience, your understanding while I was (and still am) adapting to this new field, for everything you provided me for my thesis, and for the welcome in your labs. Your advice on research has been priceless and definitely strengthened my will to continue in this direction.

I would like to thank Prof. Andrea Rinaldo, Dr. Enrico Bertuzzo, and the ECHO laboratory at EPFL for having given me the opportunity to do this project in the United States, and for contributing to several aspects of the good progress and organisation of my master thesis.

There are many things I might forget from my studies and little adventures, but I will always remember the warm welcome and help I received from my friends that made me discover the culture of the Americas. Among them I would like to mention my labmate Pamela, for always answering my myriad of (sometimes obvious) questions, advising me in my research and more, and preventing me from working too much. My other labmate, Mauricio, for our long discussions when no one was here to keep an eye on us in the lab, for having ruined my romanticism, and for also advising me in my research and more. Dea, for your always positive attitude that made me happy every time I saw you, and obviously for welcoming me in "the shelter". Alex, "who would have known that on my way to a conference I would end up" discovering that I have such a great (almost-) third labmate. Jordan, for coming to eat with me when I was abandoned in the lab, and for valorizing my unrecognised "poetry". Felicia for teaching me the contrasting aspects of the New Yorkese salsa and the negative binomial distribution. Thanks Leo for the psychological support, and making your project sound sexy and attractive (and I can tell now that it is) in order to make me come here. All of you contributed to the quality of this thesis and much more. Finally thank you to all the other people that I couldn't fit in this already-too-long paragraph.

Most importantly, a special thanks to my family. Words cannot express how grateful I am to my mother, and father, for having always believed in me, for all of the sacrifices that you've made on my behalf, giving me more than I could have ever dreamt of during my studies and life. Your guidance, advices, and encouragements are what brought me thus far and will continue inspiring my future. May I be able, one day, to give as much as I received from you.

Abstract

An explanation for the spatial variability of seasonal cholera patterns has remained an unresolved problem in tropical medicine. No simple and unified theory based on local climate variables has been formulated, leaving our understanding of seasonal variations of cholera outbreaks in different regions of the world incomplete. A mechanistic model for the Bengal region, which encompasses the variety of seasonal patterns worldwide, may provide a unique opportunity to gain insights on the conditions and factors responsible for endemicity around the globe, and therefore, to also revise our understanding of the ecology of *Vibrio cholerae*.

Through the analysis of a unique historical dataset, we propose the first mechanistic, rainfall-driven, SIR-based stochastic model we are aware of for the population dynamics of cholera, capable of capturing the full range of seasonal patterns in this large estuarine region. Parameter inference was implemented via new statistical methods that allow the computation of maximum-likelihood estimates for partially observed Markov processes through sequential Monte-Carlo.

The results indicate that the hydrological regime is a decisive driver determining the seasonal dynamics of cholera. It was found that rainfall and longer water residence times tend to buffer the propagation of the disease in wet regions due to a dilution effect, while also enhancing cholera incidence in dry regions. This indicates that overall water levels matter and appear to determine whether the seasonality is unimodal or bimodal, as well as whether it is pre-, post-, or in-phase with the monsoon. We present evidence that the environmental reservoir is responsible for the persistence of the disease, and therefore its endemicity.

Given the undeniable interplay between the seasonality of cholera and the environment, a deeper understanding of the underlying mechanisms could allow for the better management and planning of public health policies with respect to climate. In terms of disease prevention and mitigation strategies this is of paramount importance today, as changes in the population dynamics of infectious diseases are expected in response to fast anthropogenic climate change.

Table of contents

1 | Introduction

"In this village of my dreams the villager will not be dull he will be all awareness. He will not live like an animal in filth and darkness. Men and women will live in freedom, prepared to face the whole world. There will be no plague, no cholera and no smallpox."

Mahatma Gandhi,
Letter to Jawaharlal Nehru, 1939

Mahatma Gandhi once said that in the village of his dreams "there will be no plague, no cholera and no smallpox", thereby showing the predominance of these three diseases during the era of British India. Nevertheless the intriguing history of cholera, the disease at the core of this study, finds its origins much earlier in ancient times, when traces of its appearance emerged in old Sanskrit texts from the age of the *Sushruta Samihita* around 500-400 BC [1].

The scourge of the Poor

Assessed as the second leading cause of mortality among children under 5 years old and responsible for 20% of their deaths [2], diarrhoeal diseases are preventable through adapted sanitary conditions, education and hygiene [3]. Indeed, despite the fact that cholera is now one of the most famous diseases whose mechanisms and treatments are well known, it still causes despair worldwide, and does so repeatedly in the estuary of the Ganges, its native habitat. According to the *World Health Organisation*, cholera cases are nowadays estimated at around 3 to 5 million every year, taking yearly the life of 100,000 to 120,000 people [4]. A recent testimony of this explosive pattern and its unpredictable appearances is the outbreak of Haiti in 2010 which caused 7,550 deaths and 685,000 reported infections [5].

Understanding an ancient disease

Qualified as "the most terrible outbreak of cholera which ever occurred in this kingdom" by John Snow, the cholera blast that occurred in London in 1854 triggered research on the geography of epidemiology and the dynamics of infectious diseases. Because of the challenges posed by numerous human and socio-economic factors and a lack of quantitative analyses, the interplay between diseases dynamics and climatic factors has remained unresolved. However thanks to the rapid growth of climatic records, the increasing availability of data and concerns related to climate changes, the interplay has draw more and more attention [6]. In this regard Pascual et al. [6] define two major protagonists with conflicting views on the dominant drivers behind cholera's epidemiological patterns: the "localists", supporting the role of the environment and an environmental reservoir in transmission, and the "contagionists", emphasizing human-to-human transmission and sanitary conditions. Climate forecasting has now entered this research field, as understanding the role of environmental forcing on infectious diseases could henceforth contribute to their mitigation and prevention with tremendous human benefits worldwide.

1.1 Objectives and motivation

An explanation for the diverse seasonal patterns of cholera outbreaks in endemic areas has remained elusive [6]. Previous studies addressing the role of climate drivers in disease dynamics have focused on interannual variability and modelled seasonality as given [7]. The few proposing explanations for seasonality itself have relied on complex environmental interactions that vary with spatial location (involving regional hydrological models [8], river discharge, sea surface temperature, and plankton blooms). Thus, no simple and unified theory based on local climate variables has been considered [9], leaving our understanding of seasonal variations of cholera outbreaks in different regions of the world incomplete.

The Bengal region: an endemic area

Bangladesh and North-East India are an endemic region, meaning that they suffer from the persistence of the pathogen, in this case the bacterium *Vibrio cholerae*. Hence, infections occur recursively every year in this region considered to be the native habitat of the Classical cholera biotype. Here, the features of high population density, seasonal hydroclimatology, floodplain geography and coastal ecology contribute to make this region particularly vulnerable to periodic outbreaks [10].

One region, several dynamics

An endemic region exhibits both the presence of marked seasonal and interannual patterns of infection [11], and the Bengal region encompasses in this sense the most heterogeneous and widest diversity of endemic cholera dynamics worldwide. Indeed, whereas most of the areas subject to endemicity such as sub-Saharan Africa [12], southern Africa [13], Latin America [14], South-East Asia [15], show a single annual peak, North-East India and Bangladesh can also exhibit a double peak. This bimodal distribution in the frequency of cases is the most commonly observed pattern in the coastal areas of the Bay of Bengal [16], with infections pre- and post-monsoon (spring and fall). By contrast, upstream Indian states, such as Bihar, are subject to a single peak during the rainy seasons [17] and North-Eastern provinces (e.g. Assam) are almost epidemic with more irregular patterns than recurrent seasonality (chapter 3).

The aim of this thesis is to contribute to the understanding of the dynamics of cholera in one of the world's poorest and most densely populated regions. This study will do so by specifically asking the following question:

Can a model driven by local rainfall and temperature explain the wide range of cholera seasonal patterns in the Bengal region? In other words, can the variety of seasonal patterns arise from a general process whose different manifestations reflect the interaction of local climate variables and local hydrology with the population dynamics of the disease?

A rainfall-temperature driven model

Although several studies have already addressed cholera in Bengal (e.g. by Pascual et al. [11][17] and King et al. [7]), the model presented here is as far as I know the first attempt to consider only rainfall and temperature to explain the full range of patterns in cholera seasonality, with a simple and fully mechanistic approach (section 3.2.1). Indeed, only a few studies have tried to explain how single-peak environmental drivers (rainfall, temperature) can create a double annual peak in cholera prevalence [18], and for this reason, cholera dynamics in Bangladesh remain unclear. A better understanding of these dynamics has considerable implications: if a rainfall-driven model can explain this wide range of patterns using local climate variables in a mechanistic way, interpretation of the environmental and epidemic parameters could provide insights into the causes of endemicity. Moreover, because the Bengal region exhibits this diversity of dynamics

(fig. 4.2), it provides a unique opportunity to gain insights on the conditions and factors responsible for endemicity around the globe.

Mathematical models as mitigation tools

Mathematical models are important tools for understanding climatic influences in the context of the population dynamics of infectious diseases [6]. They allow the combination of the aforementioned "localists" and "contagionists" perspectives, and their relevance is evident in the scope of understanding host-parasite interactions, the timing and causal relationships of seasonality [19][20] in order to develop climate-based early warning systems for environmentally-driven infectious diseases [11]. Moreover given the wide range of conditions under which *Vibrio cholerae* can thrive, its complete eradication is very unlikely; hence the importance of understanding its close relationship with the environment in order to mitigate it [6]. Of particular interest in these last decades is the response of infectious disease dynamics to anthropogenic climate change [19] as those changes can alter various processes such as parasite spread [21], host behavior, immune function, etc.

This leads to the second question this thesis will address:

Under which conditions do the different seasonal patterns emerge in this endemic area?

Through the analysis of a unique historical dataset containing 50 years of monthly meteorological, demographic and epidemiological records, we propose a process-based model for the population dynamics of cholera. This model is driven only by local rainfall and temperature but it is able to capture the full range of seasonal patterns in this large estuarine region. This analysis is conducted using a mechanistic, partially observed, nonlinear, stochastic dynamical SIR-based model (chapter 3), which is confronted to the 50 years of records in the district of Bengal of former British India (now Bangladesh and the Indian states of West Bengal, Bihar and Assam). The relevance of rainfall as a main climatic driver is implied by the waterborne nature and transmission of the bacterium, and its link to the monsoons [17], whereas temperature is an obvious factor because of its direct effect on the ecology of microbial populations [16]. The fitting procedure is implemented via iterated filtering [22], a method for computing maximum-likelihood estimates for partially observed Markov processes through sequential Monte-Carlo [23] (chapters 3.3.1 and 3.3.2).

1.2 Thesis structure and chapter description

A journey in 7 chapters

This thesis contains 7 chapters including this introduction. It is then articulated according to the following structure:

- Chapter 2, *Context*, contains an overview of the disease and the area under study, namely North-East India and Bangladesh. The geography, hydrology and climate of this endemic area are briefly described as well as *Vibrio cholerae*'s ecology and survival mechanisms.
- Chapter 3, *Litterature review and Technical Background*, provides a short summary of the previous work on infectious disease dynamics and on cholera in particular, in relation to the goals of this thesis. The chapter is divided into two different parts, a first one that reviews the literature and models, and a second one, that details the recently developed methods used to fit the model to the data.
- Chapter 4, *Cholera in Bengal - An Historical dataset*, consists of a brief description of the extensive dataset used in this study: the records of 50 years of meteorological

variables and cholera deaths for British India. The main temporal patterns, in particular in the seasonality, are examined.

- Chapter 5, *Modelling and Methodology*, brings to the stage the model at the core of this study and the numerical implementation for its simulation and parameterization.
- Chapter 6, *Results and Discussion*, is an overview of the results obtained for various regions in the study area. The seasonality, model parameters and interannual variation are described and followed by a discussion. Finally it suggests the future directions for the further understanding and of cholera dynamics in endemic areas.
- Chapter 7, *Conclusion*, the last chapter, wraps up this thesis by presenting the theoretical and practical implications of this work.

2 | Context

2.1 Cholera - *Vibrio cholerae*

Cholera is a waterborne enteric disease characterized by severe watery diarrhea caused by the gram-negative bacteria *Vibrio cholerae*. The bacteria colonizes the small intestine and after infection, without immediate treatment, the afflicted individual can die of dehydration within hours [20]. Present in over 70 countries, the disease is now in its seventh pandemic.

Two biotypes

Nowadays around 200 serotypes have been identified based on the O antigen of *Vibrio cholerae*'s lipopolysaccharide. However among those 200 only O1, and more recently O139, are toxigenic strains producing virulence factors [24]. It is the O1 serotype that contains the two famous biotypes: Classical and El Tor. Being nowadays the dominant biotype, El Tor pandemics appeared only in the 1960's, it is the Classical strain which is responsible of the 19th-20th century pandemic covered in this study. Although the two biotypes have an epidemic potential, El Tor shows less severe infections, leading to a higher asymptomatic to symptomatic ratio and a lower pathogenicity [25]. This asymptomatic to symptomatic ratio is of significant importance as only 1 to 20% of the infected individuals show sever symptoms [7] but all of them contribute to the disease spread. The growth conditions of the two biotypes are similar, even though laboratory studies suggest more specific conditions and a higher maximal activity (37°C, corresponding to the human intestine and spring-summer months in the Bay of Bengal) for El Tor compared to Classical (30°C) [26].

Why is this pathogen here?

In the recent years, 4 mechanisms have been proposed in the literature in order to explain the presence of *Vibrio cholerae* in endemic areas and its recursive fadeouts. According to the London School of Hygiene and Tropical Medicine those mechanisms are [27]:

- the maintenance of the O1 biotype in non-human populations
- the maintenance in chronic carriers not necessarily excreting the organism
- the maintenance by asymptomatic infected humans
- the maintenance of the pathogen in an aquatic reservoir

It is nowadays the last point, an aquatic reservoir maintaining endemicity and the relevance of primary transmission, that attracts the most interest and is actively discussed by the scientific community [6] [20] [28] [11] [29] [16] [10] [30] [15] [31]. Hence its integration as one of the 7 compartments of the SIR model (see chapter 5).

The spread of the pathogen in the environment is mainly governed by 4 environmental variables:

- Temperature
- pH
- Salinity
- River discharge

Temperature

As any microbial fauna, *Vibrio cholerae* is subject to temperature restrictions. Pascual et al. [32] mention that high temperatures cause a warming of the continent and its water bodies, which are the bacterium environmental reservoir through which transmission occurs. More importantly temperature can change human behaviour and contact rate with contaminated water. For example, the highest temperatures, coincide with the dryer period which occurs in the pre-monsoon months, and can in turn result in a higher contact rate of the population with the restricted availability of water.

Sea Surface Temperature (SST) has also been proposed as "a major factor driving cholera outbreaks" [47]. Bouma and Pascual [28] found a significant relationship between SST in the Bay of Bengal and the spring peak in coastal districts for the earlier 20th century.

pH and salinity

Vibrio cholerae is a bacterium adapted to alkaline environments. Studies in acidic surface water of tropical rainforests in South America confirmed the absence of the pathogen and thus the relevance of the pH [6] [33]. Saline water bodies are also part of the bacterium optimal growth conditions. Studies on the effect of temperature and growth [34] showed that the dependency on salinity in unequivocal : at 10 °C extended survival (up to 42 days) was observed with salinity levels of 25 ‰versus 4 days for 5 ‰.

River discharge

Although deserving further attention, river discharge is another factor presumably responsible for cholera outbreaks. Jutla et al. [35] suggest that it is river discharge, and not SST, which is responsible for the positive correlation between cholera incidence and SST in the Bay of Bengal. The higher summer river discharge conveys important nutrients loads in the Bay, resulting in phytoplankton blooms then followed by zooplankton blooms primarily by copepods on which *Vibrio cholerae* forms biofilms. These organisms would then be carried inland during the dry and low-discharge period with the intrusion of salt water, creating optimal survival conditions for the cholera pathogen [33]. This would in turn increases human exposure in areas in the proximity to water bodies.

Moreover it is worth mentioning that high discharge leads to flooding, which given the poor sanitary infrastructure can increase the pathogen's spread, human concentration, and the susceptibility of people to disease in crowded and stressful conditions, favouring secondary transmission route.

Transmission modes

The pathogen has two stages in its life cycle, a human and an environmental one, with two main transmission modes. Primary transmission occurs from the environmental reservoir to the host whereas secondary transmission occurs from human to another human directly or via the environment[1] (e.g. fecally contaminated water or food).

[1] In the secondary transmission the force of infection is linked to the cases, whereas primary transmission often think of the pathogen living in the environment and from there causing exposure.

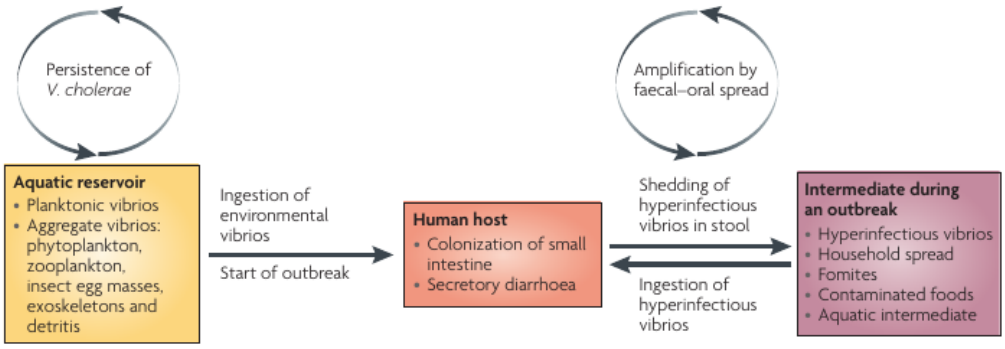


Figure 2.1: Life cycle of pathogenic *Vibrio cholerae* [24].

It seems thereby clear that the Bengal region with its brackish waters and estuaries provides for *Vibrio cholerae* and ideal habitat to thrive outside the human host. Moreover given *V. cholerae* ability to enter a "dormant" stage, postulated as such because the bacterium is present but non-culturable during certain times of the years, an extended viable period is possible under unfavourable environmental conditions. Because of this ability to survive in a wide range of conditions, the eradication of cholera is unlikely. From this perspective, the further understanding of the disease is paramount in order to mitigate its ravaging effects.

2.2 The Bengal region

The Jewel in the British crown

Once known as *The Jewel in the British crown*, British India was in the 19th century one of the most valuable provinces of the British Empire thanks to its important manpower and trade benefits. Initiated by Jawaharlal Nehru in the 1920's-1930's, the independence movement succeeded in releasing India from its Occidental leverage in 1947. This duly acquired freedom led the country to face massive demographic, societal and political changes such as the Indo-Pakistani War, the famous Green Revolution, the division of the Bengal region and several religious cleavages. Thus at the departure of the British East India Company, the Bengal region was divided between India and Pakistan based on religion. In 1971, the Pakistani part (known then as East Bengal) claimed its independence and became the country of Bangladesh we know today.

2.2.1 Geography, topography and hydrology

Area under study

The Bengal region is thus an eastern region of the Indian subcontinent and is composed of the Indian state of West Bengal and the nation of Bangladesh. It is a low-lying Ganges Delta, the world's largest delta, and hosts around 245 million people over an area of 232,752 km^2, making it one of world's most densely populated region [36]. Dhaka, the capital and biggest city, shelters around 15 million people at the confluence of the three biggest rivers : the Ganges, the Brahmaputra and the Meghna. The studied area of the present thesis includes as well the Indian states of Assam (north-east), Bihar (north-west), Meghalaya and Tripura (east) (figure 6.1).

Topography

Bangladesh is often referred to as an estuary; it is subject to low terrain elevation, brackish water, inland rivers and coastal proximity. Figure 2.3 illustrates this status, showing that most of the country has an altitude of less than 10 meters above sea level.

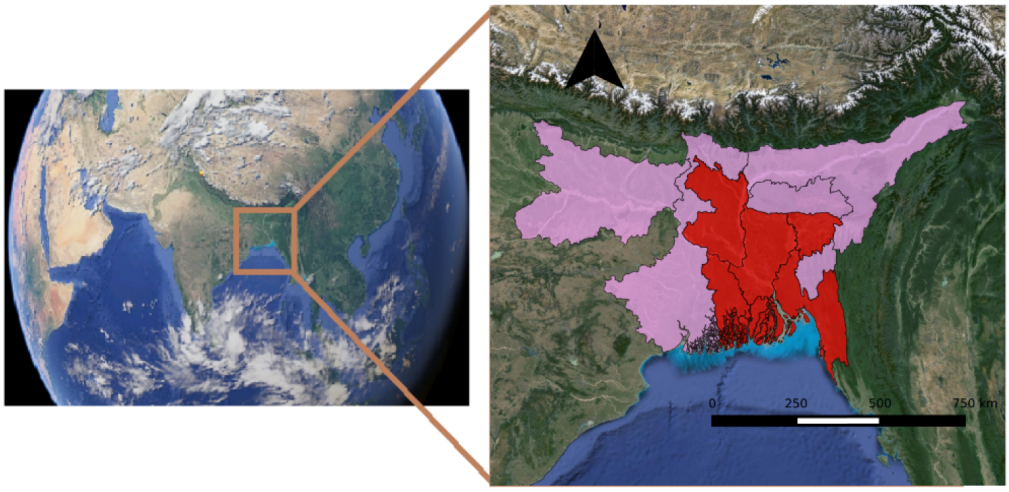


Figure 2.2: The Bengal region contains 5 states of North-East India, in purplish pink, and Bangladesh, red (QuantumGIS and the Openlayer plugin, district data from [37]).

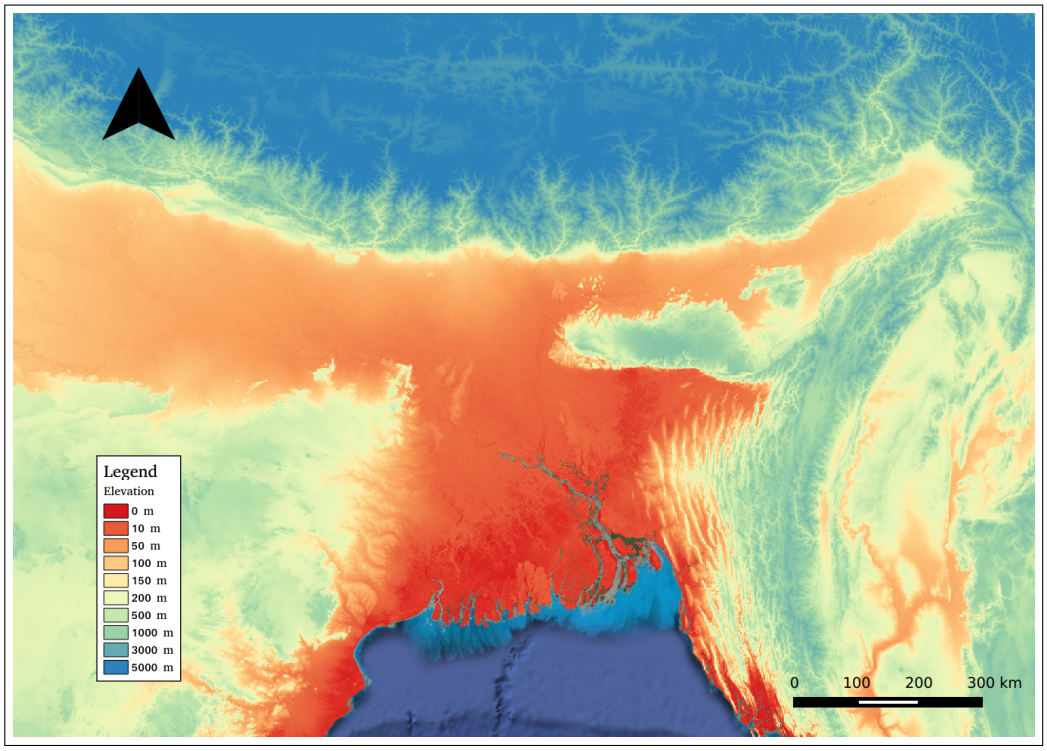


Figure 2.3: Elevation map of the Bengal region (QuantumGIS, DEM data from [37]).

Figure 2.3 displays as well an important feature of the area, the sharp Himalayan wall. This mountain range is responsible for the Indian monsoon as the high moisture of the summer south-north winds suddenly encounters the mountains creating the important summer precipitations by orographic lift.

Hydrological network

This part of the Indian subcontinent exhibits a complex hydrological network (fig. 2.4). Two main rivers whose source is found in the Himalayas are responsible for the high discharge observed in the delta of the Bay of Bengal: the Ganges (north-west to south-east) and the Brahmaputra (north-east to south-west in figure 2.4). The Meghna river is also a significant contributor to this hydrological regime.

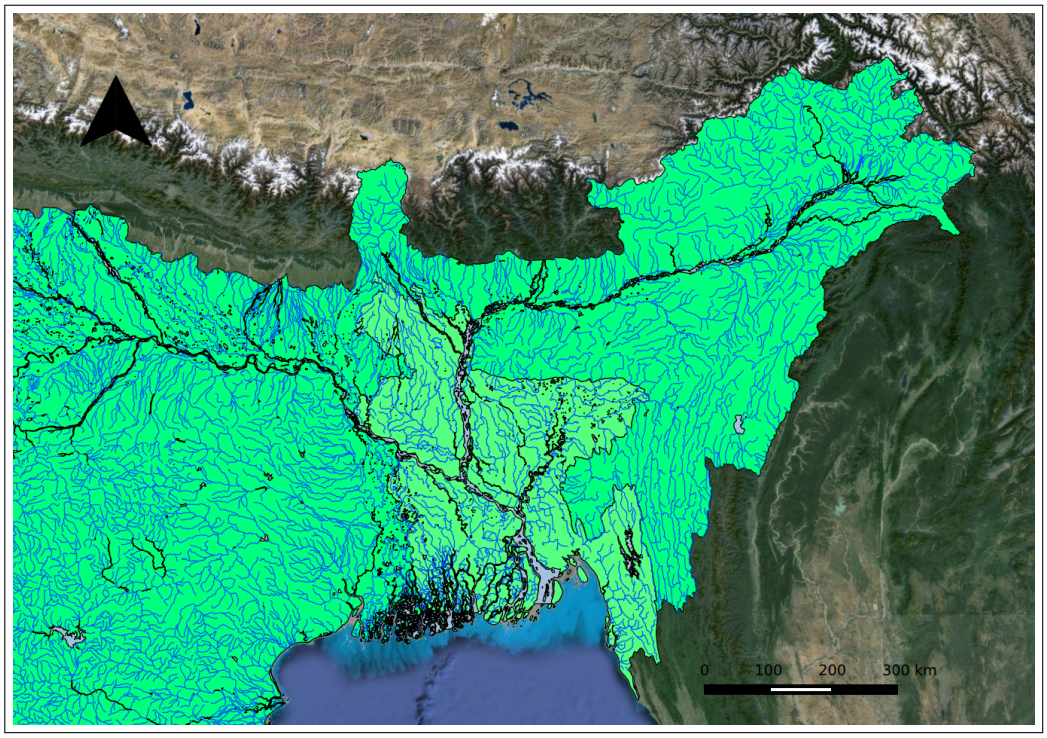


Figure 2.4: Hydrological network of North-East India and Bangladesh (QuantumGIS and the OpenLayers, hydrologic data from [37]).

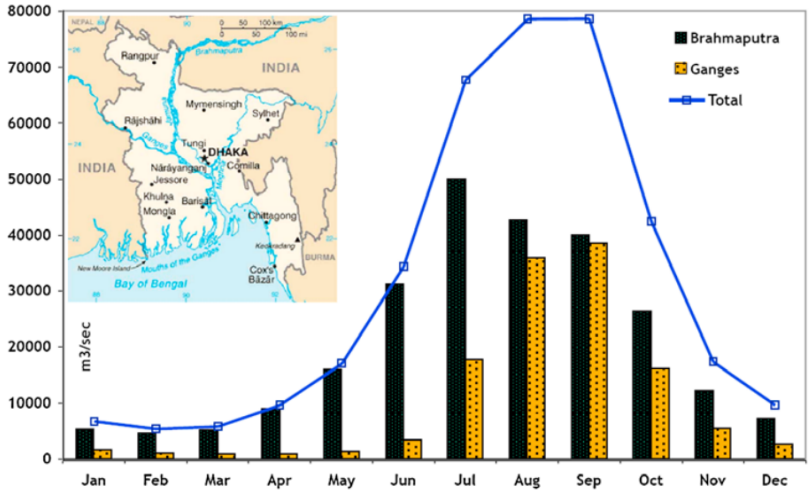


Figure 2.5: Monthly discharge of the Ganges and Brahmaputra [10].

In terms of discharge the Ganges and Brahmaputra are under a seasonal regime

showed in figure 2.5. This discharge is the result of snow melt in the Himalayas and runoff on the low-lying inland regions of Bengal. A maximum discharge is observed in the late summer months, reaching an impressive peak of 80,000 m^3/s in August and September. Those high discharges are often linked to important flooding in the delta and upstream regions, interconnecting water bodies across the landscape and potentially spreading the pathogen from infected reservoirs. Moreover even though the salinity decreases, the high nutrient loads brought to the Bay of Bengal contribute significantly to plankton blooms, which have been hypothesized by some authors to drive the fall-winter peak in coastal regions.

2.2.2 Population and sanitation [2]

Sanitation in Bangladesh

Sanitation is still one of the main concerns of the new government of Bangladesh, especially because it is directly linked to a variety of health problems. "From 2003 to 2006, the Government of Bangladesh (GoB) scaled up efforts to address unsanitary household practices through a national sanitation campaign that engaged multiple levels of government. The government's goals were to achieve 100% sanitation coverage and stop open defecation in rural areas by 2010" [39]. In October 2011 a study of the *Water and Sanitation Program* [40] done in 36 districts of Bangladesh showed that "89.5% of sample households own or share a latrine that safely confines feces. Of the remaining 10.5% of the households, 2.5% do not have any latrine, 5.5% have a hanging latrine or facility that drains directly into the environment and 2.5% use an open pit without a slab". The important point here is that out of the 89.5% of households with latrines, only 37% met the criteria for a "hygienic" latrine according to the GoB definition.

Regarding access to improved water sources, it is estimated as relatively high for a low-income country, reaching a coverage up to 98%. However, in the mid 90's, it was discovered that a high percentage of the ground water was naturally contaminated with arsenic, leading some people not to use the existing wells anymore and going back to the consumption of surface water, which can be easily contaminated due to the poor sanitary condition aforementioned.

2.2.3 Climate[3]

Climate and seasons

Straddling the Tropic of Cancer, Bengali climate is tropical and indeed warm and humid. This climate is mostly influenced by monsoon and partially by pre- and post-monsoon circulation. The monsoon has its onset during the first week of June and withdraws in the first week of October, but it could vary from year to year [41].

Ahsan Uddin Ahmed explains in his paper *Bangladesh Climate Change Impacts and Vulnerability* that there are four prominent seasons, namely winter (December to February), pre-monsoon (March to May), monsoon (June to early-October) and post-monsoon (late-October to November). The general characteristics of the seasons are the following [41] :

- Winter is relatively cooler and drier, with the average temperature ranging from a minimum of 7.2 to 12.8 °C to a maximum of 23.9 to 31.1 °C. There is a northward thermal gradient in winter, meaning that temperature in the southern districts are 5 °C warmer than in the northern districts.

[2] Subsections 2.2.2 and 2.2.3 were written by L. Evéquoz in his master thesis [38].
[3] See footnote 1.

- Pre-monsoon is rather hot with an average maximum of 36.7 °C, meaning there is a very high rate of evaporation or important drought and thus erratic but occasional heavy rainfall from March to June. The maximum temperatures are observed in April at the beginning of pre-monsoon season. In pre-monsoon season the mean temperature gradient is oriented southwest - northeast with the warmer zone in the southwest and the cooler zone in the northeast.
- Monsoon is both hot and humid and brings heavy torrential rainfall throughout the season. About four-fifths of the mean annual rainfall occurs during the monsoon. The mean monsoon temperatures are higher in the western districts compared to those in the eastern districts.
- Post-monsoon is a short-living season characterized by the withdraw of rainfall and gradual lowering of night-time minimum temperatures.

The mean annual rainfall is about 2300 mm, but there is a wide spatial and temporal distribution. Annual rainfall ranges from 1200 mm in the extreme west to over 5000 mm in the east and north-east (MPO, 1991). Generally, the eastern parts of the country receive higher rainfall than the western parts [41].

3 Literature review and technical background

3.1 Literature review

3.1.1 Endemic cholera throughout the World

It is in the 19th century that cholera started to spread across the globe from its homeland, the estuary of the Ganges triggering the 6 pandemics that killed millions of people across every continent. Since that time cohorts of scientists and doctors began to fight this disease trying to understand its mechanisms, and nowadays a plethora of scientific papers and discoveries exist about cholera. Because of its endemic nature and as it is believed to be *Vibrio cholerae*'s native habitat, the Bengal region has been widely studied and documented. Thus, without pretending to completeness, this section will intent to grasp and summary the main concepts and ideas discussed in this abundant literature, especially with respect to South Asia.

The variability of endemicity

Although during the era of British India few places were endemic, it is worth mentioning that nowadays several countries suffer from the recursive presence of the pathogen, which became a worldwide concern. Emch et al. [9] worked on inventorying the global cholera patterns worldwide. These endemic dynamics are now observed in the following countries:

- **Bangladesh**, seasonal cholera epidemic in April and November-January
- **Kolkata**, April to June
- **Pakistan**, from November to January and April to May
- **South America**, January to February (summer months)
- **Amazon region** in Brazil, end of the rainy season
- **Eastern Africa**, Djibouti, Kenya, Somalia, Uganda and Tanzania, during/after the summer rains
- **Mozambique**, December to May (rainy months)
- **South Africa**, end of January to mid-march
- **US** Gulf Coast and Australia, warmer months and late summer in the US [16]
- **Europe**, some cases reported during the summer and early fall [16]

It can thereby be noted that a wide range of endemic patterns exists across the globe and this diversity is fully represented in the Bengal region, making of it a unique area of study. Chapter 3.1.2 presents a detailed overview of the dynamics observed in the Indian provinces of Assam, Bihar, Meghalaya, Tripura and in Bangladesh.

3.1.2 Reasons for endemicity in North-East India and Bangladesh

Although endemicity refers to a periodical pattern, it can be highly dynamic and vary dramatically in intensity, frequency and duration [9] around the world. Indeed, even in Bengal, the seasonality exhibit strong geographic variations. Moreover, as can be seen in chapter 6 the interannual variation of the cholera outbreaks, for example, is as well subject to significant changes. One can thus wonder what drives these dynamics, why is it seasonal in some areas and completely epidemic in some others. The reasons are both environmental and demographic.

While chapter 2.1 already described the main environmental variables responsible for the appearances of the pathogen, the following sub-sections will focus on the mechanisms identified by the literature that are relevant for a global understanding of cholera dynamics.

Environmental drivers

Mechanisms of seasonality

Colwell et al., in *The Ecology of Vibrio cholerae* [42] report that until 1970-80, "*Vibrio cholerae* was believed to be highly host-adapted and incapable of surviving longer that a few hours outside the human intestine". Ironically, Robert Koch himself, the scientist who identified the bacterium, believed that the bacillus was able to survive and flourish in the environment. And indeed, cholera being a waterborne disease, its relationship with water bodies and the environment is quite obvious and most of the literature now agrees that environmental factors affect the distribution and frequency of the disease.

The contribution of the environmental framework is most likely the reason why others deltas of South-East Asia do not maintain cholera transmission even with similar population densities and sanitary conditions than the delta of the Ganges [6]. Indeed, the bacterium belongs to the tropical realm, emancipates and flourishes in estuaries, laminar rivers, seas and coastal areas, North-East India and Bangladesh provide to *Vibrio cholerae* all these conditions in a single regional place.

As mentioned, the literature clearly identified several seasonal patterns in the Bengal region. However one the most discussed dynamic remains the double annual peak observed in Bangladesh and especially in Dhaka, its capital. This is partly due to the fact that this pattern is quite unique as most of the endemic areas around the world are subject to a single peak. Moreover due to the numerous climatic, demographic, and epidemiological records available since the colonial era, it provides one of the richest dataset worldwide.

The double peak

In their paper *Hydroclimatic influences on seasonal and spatial cholera transmission cycles*, Akanda et al. [18] found that "cholera outbreaks in the Bengal Delta region are propagated from the coastal to the inland areas and from spring to fall by two distinctly different transmission cycles, pre-monsoon and post-monsoon, influenced by coastal and terrestrial hydroclimatic processes".

The first outbreak, occurs in spring from March to May. The authors have hypothesized that this first peak is mainly modulated by coastal hydroclimatic conditions (SST, salinity, plankton abundance) by opposition to the second one more dictated by inland processes. Thereby, during the low river discharges of the dry spring season, sea water penetrates inland (Rahman et al. [43] report intrusions of sea water up to 100 km inland during very dry years) where it triggers the first annual cycle of outbreaks. Ruiz et al. [17] have proposed to assign this first infection cycle to a temperature increase and a dry environment, which increases the pathogen concentration and the human contact rate with it due to scarce water availability and high temperatures (as the water is rare it will be consumed more and less wasted hence increasing the contact rate). This first peak is thus characterized by a strong primary transmission.

The importance of the hydrological regime

In Bangladesh, the important summer rains have shown to induce a dilution effect [44], presumably lowering the incidence of cholera. Akanda et al. hypothesize that the peak streamflow observed in June creates important inundations spreading the pathogen across the landscape and creating a large-scale contamination. Although those important discharges lower salinity levels and pH[1], they are also loaded with nutrients, which end up favoring plankton flourishing and bacteria development. These events set the perfect stage for the second annual cycle of infection in fall once the rain and discharges decrease. Finally the decline in cholera infection observed in January and February is suspected to be temperature related [6]. Moreover, thanks to its ability to enter a nonculturable state, the bacteria can survive most of the year, even during the winter, in aquatic reservoir, ready to strike again the following year [45][46].

In another of their articles [10], Akanda and Jutla have proposed that "low flow in the Brahmaputra and the Ganges during spring is associated with the first outbreaks of cholera in Bangladesh" and that "peak flood volumes and extent of flood-affected areas during monsoon are responsible for autumn cholera outbreaks".

The single peak

In *Cholera and climate: Revisiting the quantitative evidence* [6], Pascual and Bouma study the role of climate on cholera seasonality within 24 districts of the former British India (1989-1940). They emphasize that a monsoon-associated pattern (a single annual peak) was also observed in Bengal. This completely opposite pattern (fig. 3.1) is found in the dryer parts of Bengal, "indicating that overall water levels matter and appear to determine whether the effect of rainfall is positive or negative", thereby showing once again the importance of the hydrological regime and water reservoir with respect to cholera in the area.

[1] Recall: *Vibrio cholerae* likes alkaline pH

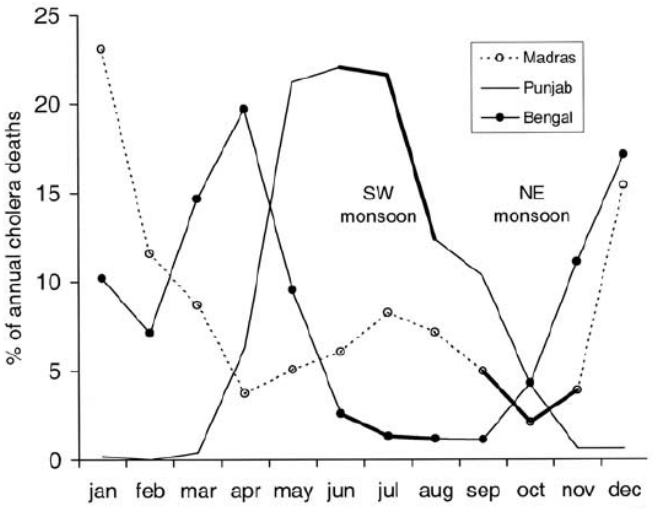


Figure 3.1: Variability of seasonal patterns. Double peak in Madras and Bengal versus single one in Punjab (more epidemic province). Bold lines correspond to the rainy season [6].

Studies about plankton

Numerous studies mention the importance of plankton in the dynamics of the disease. Torres Codeco's findings even suggest jumping from a "two factor system" (human-pathogen) to a "three or four factor system" ((virus-)bacteria-plankton-human) [20]. Colwell [47] explains the cholera outbreaks in Peru and Bangladesh by a link with the seasonal zooplankton growth. Indeed it is hypothesized that the pathogen could attach to copepods and cyannobacteria as they are feeding from zooplankton, which in turn are feeding from phytoplankton, thus increasing *Vibrio cholerae*'s survival.

In another paper by Bouma et al., *Seasonal and interannual cycles of endemic cholera in Bengal 1891-1940 in relation to climate and geography* [28], the spring outbreak was shown to be significantly correlated with the sea surface temperatures in the Bay of Bengal. These SST have, in turn, been associated to plankton bloom as a possible explanation for cholera outbreaks by Colwell.

Finally it is worth mentioning the study by Jutla et al. [35], answering to the question "phytoplankton abundance is inversely related to sea surface temperature so why a positive SST-phytoplankton relationship exists in the Bay of Bengal?". The authors have hypothesized that this relationship is valid only for the coastal regions because of the nutrient brought by the high river discharges occurring at the same time than the SST increase.

Vibriophages

Few mentioned the importance of vibriophages in the regulation of endemic cholera. One interesting *PNAS* paper by Faruque et al. [48] found that "the presence of bacterial viruses acting on V. cholerae O1 or O139 (cholera phages or vibriophages) inversely correlates with the occurrence of viable V. cholerae in the aquatic environment and the number of locally reported cholera cases". According to the authors, "cholera phages can influence cholera seasonality". Such a concept is particularly important when the bacterium is introduced to a new cholera-free area, which will lack those natural predators to regulate epidemics and might thereby be subject to an explosive outbreak.

El Niño

El Niño is the most important phenomenon of interannual climate variability on a global scale and its influence on various ecological processes are numerous. El Niño

Southern Oscillation (ENSO) influence the climate of many places of the world, and the Indian Ocean is no exception to this leverage. After the warming of the Pacific, changes in cloud cover, evaporation, and increased heat flux can be observed a few months later in the Bay of Bengal, thus linking general climate to local variables impacting cholera.

In an article published in *Science* [32], Pascual et al. assess the role of ENSO in the dynamics of the disease. Using nonlinear time series analysis and the Niño3.4 index (an index based on the SST anomaly in a region of the equatorial Pacific), they found that in Bangladesh, cholera dynamics "are consistent with a remote forcing by ENSO". The incidence of the disease increase with negatives values of the SST anomaly, while it decreases with higher positive values. Moreover a dominant period of 3.7 years has been identified in the cholera cases between 1980 and 1998, which matches exactly the one of Niño3.4. Thanks to Singular Spectrum Analysis (SSA), a method used to decompose time series in order to extract the interannual variation and seasonality, they found that "this frequency accounts for 46% of the variance in the interannual variability of cholera".

Finally, Rodo et al. in a paper from *PNAS* [49] used again SSA to identify that, in Dhaka, "a strong and consistent signature of ENSO is apparent in the last two decades". ENSO accounted for over 70% of the cholera variance in Dhaka from 1980 to 2001 and during time windows of local maxima, thereby suggesting a nonstationary link between the disease and ENSO (fig. 3.2). For the historical time series however, this relationship was weaker and highly irregular.

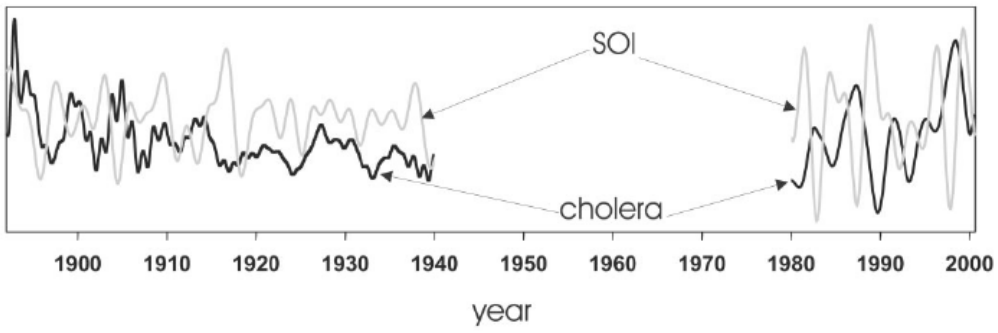


Figure 3.2: Reconstruction of the first 4 principal components of the SSA for cholera and SOI (Southern Oscillation Index) [49]

As mentioned by Emch et al. [9], it seems henceforth clear that cholera is a disease linked to climate, under the influence of both micro- and macro-level environmental parameters.

Demographic drivers

A burden linked to socio-economic factors

Also known as *The Scourge of the Poor*, cholera is obviously a disease bounded to socioeconomic and demographic drivers. Studies have shown that with good sanitation, some cases of primary transmission could occasionally occur but no secondary transmission is possible [20]. In poor communities, endemicity could be maintained even without an environmental reservoir, for example due to an exposure to untreated sewage. Heavy rainfall during the monsoon can easily cause a breakdown of the sanitary infrastructures in developing countries, thus increasing the contact with the pathogen and triggering an outbreak. Ruiz-Moreno [17] indeed found an "enhancing effect on secondary transmission

during extreme rainfall events".

In *Nature*, King et al. [7] found that the asymptomatic to symptomatic ratio is "far higher than had been previously supposed", which is typical from endemic areas where the infections and mortality are milder. Hence those asymptomatics act as "silent shedders", and are especially important in regions with poor sanitation as they will propagate the pathogen even though one might think the conditions are safe due to only few apparent cases. Another interesting finding of the study is the analysis of the susceptible pool, which, when depleted, bring the epidemic to a halt. However as the immunity disappears after weeks to months (other studies found duration of immunities up to several years, for the outbreak in Haiti [50]), the susceptible pool is replenished and ready to suffer from the next outbreak.

Bellew et al. [51] and Altizer et al. [19], conducted interesting researches on the variation of the immune system of the human host. The first study identified 3-year cycles with cholera data for India from 1862 to 1881 and years of "drought and famines at times of peak incidence", suggesting the weakening of the host immune system due to malnutrition and thus a higher susceptibility to cholera. The latter found as well rhythms that shape the immune function (harsh weather conditions, winter, poor nutrition).

Finally Ruiz-Moreno et al. [17] link the endemicity of an area to its human density. Their analysis conducted in Madras showed that districts with a higher density have fewer cholera fade-outs compared to the ones with a low population density. Hence more stochastic patterns and thus epidemic dynamics are more prone to be observed in regions with lower human population (e.g. Northern parts of the Bengal region).

3.2 Disease modelling

3.2.1 SIR models

1920's, the first SIR model

The first compartmental epidemiological models originated from the 1920's, where they were initially used to explain the rapid growth and fade of infected individuals during epidemics. The simplest form of those models is the SIS, standing for Susceptible, Infected and Susceptible. The population is divided into those 2 compartments (S and I) and move from one to the other according to a system of differential transition equations. An SIR model is similar except that it considers the acquisition of an immunity after an infection, hence the addition of a new state accounting for the Recovered. Its simplest form, known as the Kermack-McKendrick model (in reference to their creators), can be visualized in the following equations:

$$\frac{dS}{dt} = -\beta SI \tag{3.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{3.2}$$

$$\frac{dR}{dt} = \gamma I \tag{3.3}$$

Where β is the infection rate, γ the inverse of the duration of infection and $-\beta I$ the force of infection. The model considers a lifetime immunity and constant population. The susceptible state (S) contains the healthy individuals threatened by an infection, the

infected state (I) the ones that contracted the disease, and the recovered (R) the ones that acquired a post-infection immunity.

Representing the unknown laws of nature

Nowadays those models have been widely extended, some incorporate an Exposed state, some others an Inapparent Infected state [7], the addition of a compartment for the pathogen [20], for the vectors of the pathogen [52], partial immunity [52], a water reservoir [20], etc. Those models are usually semi or fully mechanistic, meaning that they require an understanding of the natural processes as they model the dynamics from a phenomenological point of view. They are a representation of the unknown laws of nature and their parameters have therefore a scientific and interpretable meaning. An example of one of those models, which served as a springboard for this study is introduced below.

A stochastic SIR model

An SIR model that inspired significantly this study is the one developed by King et al. in a *Nature* paper published in 2008 [7]. There they study as well cholera dynamics in Bengal during the colonial period[2], however they rely on a "semi-mechanistic" approach where the seasonality is modelled as a flexible periodic function of time, which allows to overcome its complex nature. To do so new statistical methods allowing maximum likelihood inference based on stochastic dynamical partially observed models were used [3].

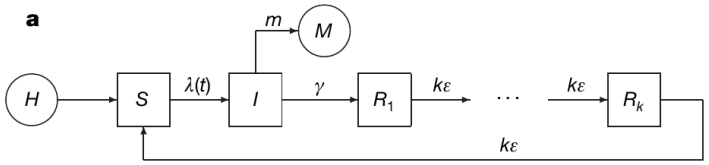


Figure 3.3: SIR model. "Births, related to the total population size H, are assumed to feed the pool of susceptibles, S. Individuals are susceptible to infection when born. Exposure to the pathogen occurs at time-dependent rate $\lambda(t)$. Infected individuals die at an excess rate m and recover at a rate γ; the time an individual spends within the I class is exponentially distributed. We assume that an individual remains immune to reinfection for a duration gamma-distributed with mean $1/\epsilon$ and variance $1/k\epsilon^2$. Once immunity has waned, an individual re-enters the susceptible pool (S). The measured variable is monthly deaths, M. Individuals in each class are subject to constant background mortality at rate 0.02 y^{-1}. The force of infection includes terms for environmental and human sources of infection and is assumed to vary seasonally" [7].

The simplest form of the models proposed by King et al. (fig. 3.3) is written as:

[2]Using the same epidemiological and demographic records than in this study.
[3]Those methods are detailed in the next sections.

$$\frac{dS}{dt} = \kappa\epsilon R_k + \delta H(t) + \frac{dH}{dt}(t) - (\lambda(t) + \delta)S \tag{3.4}$$

$$\frac{dI}{dt} = \lambda(t)S - (\gamma + m + \delta)I \tag{3.5}$$

$$\frac{dR_1}{dt} = \gamma I - (k\epsilon + \delta)R_1 \tag{3.6}$$

$$\vdots \tag{3.7}$$

$$\frac{dR_k}{dt} = k\epsilon R_{k-1} - (k\epsilon + \delta)R_k \tag{3.8}$$

$$\tag{3.9}$$

With environmental stochasticity entering through noise added to the time-varying force of infection $\lambda(t)$, modelled according to:

$$\lambda(t) = w + [\bar{\beta}e^{\beta_{trend}t}\beta_{seas}(t) + \sigma\xi(t)]\frac{I(t)}{H(t)} \tag{3.10}$$

Nature and stochasticity

Most of the processes observed in nature are often a single realisation of a myriad of possible outcomes. Considering stochasticity allows taking into account this randomness, which can be significant in the dynamics of some phenomenons. Thereby it is here introduced trough process and measurement noise. The process noise being here $\xi(t) = dW/dt(t)$, a Gaussian white noise. $1/\epsilon$ is the mean duration of immunity, $1/\gamma$ the mean duration of infection, $1/\sqrt{k}$ the coefficient of variation of the immune period, $1/\delta$ the life expectancy, m the cholera mortality rate and $H(t)$ the population size from census data. Finally the computed deaths of month n, M_n, are related to the reported ones, y_n by a conditional Gaussian distribution, $y_n \sim normal(M_n, \tau^2 M_n^2)$, with $M_n = m\int_{(n-1)/12}^{n/12} I(t)dt$.

Creating seasonality

The particularity of this model resides in its force of infection, which considers both primary and secondary transmission. Hence while w is independent from the infective pool (non-human reservoir of pathogen), $[\bar{\beta}e^{\beta_{trend}t}\beta_{seas}(t)+\sigma\xi(t)]\frac{I(t)}{H(t)}$ corresponds to the human-to-human transmission. Moreover an artificial seasonality is induced by modelling a time-varying contact rate with:

$$\log\beta_{seas}(t) = \sum_{k=0}^{5} b_k s_k(t) \tag{3.11}$$

Where $\{s_k(t)\}$ is a periodic cubic B-spline basis used to construct the flexible periodic function. In this case the seasonal component is generated with 6 splines, adding suchlike a significant number of parameters to fit but at the benefit of providing an important flexibility to the model. β_{trend} is a term allowing a long term change in time of the contact rate (e.g. due to a decrease in reported deaths) and $\bar{\beta}$ a scaling constant. Finally σ is an environmental stochasticity parameter.

3.3 Fitting methods

Properties and terminology

Once a model is selected, a common problem is to infer the parameters of the system, which are unknown. Those parameters are obviously required to estimate the unknown states from the observed ones. To do so, this study considers a method called *Maximum likelihood via Iterated Filtering*, which allows maximum likelihood estimation for partially observed nonlinear dynamical systems. However, before jumping into the technical part and description of this method, here is a brief definition of relevant mathematical terms.

The SIR model built in this study is a *mechanistic, partially observed, non-linear, Markovian, stochastic, dynamical system [23]*. Where:

- **mechanistic** has already been defined in the previous section.
- **partially observed** accounts for the fact that not every state is measured[4] (e.g. in the case of cholera only the deaths are, not the susceptible, infected, etc.).
- **non-linear** is a property shared by the physical processes of most of those systems.
- **Markovian** implies that the states of the future depend only on the states of the present, and are therefore independent from the states of the past.
- **stochastic** denotes that some randomness is introduced in the model. Two sources of stochasticity are added here : the process and measurement noise.

How to estimate those unknown quantities?

[5] Estimating unknown quantities (state variables in the SIR model) from a few measurements is widely sought and used in a variety of domains (science, engineering, economy, etc.). In many of these applications preliminary knowledge of the processes modeled is available, which allows for formulating *a priori* distributions of the unknown states[6][53]. When this is the case, Bayes' theorem provides the *a posteriori* distribution, which in turns enable us to infer the unknown quantities. However this approach becomes problematic when playing with data available sequentially, as the posterior distribution will have to be updated every time a new observation is available. A perfect example of such a situation is the positioning by GNSS[7], where the position, the unknown quantity, is inferred from the posterior distribution obtained from the received data. In the case of satellite positioning however, the system is linear and the noise distribution is often approximately Gaussian therefore, the *Kalman filter* is widely used as it provides an "analytical solution of the evolving sequence of posterior distributions" [53].

A powerful and elegant solution

The *Kalman filter* is an attractive alternative in this kind of situation, but a drawback exists: it is valid only for linear Gaussian models, which is not the case for the one considered in this thesis. Doucet et al. mention other alternatives that have been used for over 30 years but either some yield to poor results, or are too computationally expensive. It is in this context that the powerful and elegant *Sequential Monte-Carlo* methods enter into action.

[4] Those observed states are often measured with noise

[5] This paragraph is based on the introduction to sequential Monte-Carlo methods by Doucet et al. [53]

[6] When those prior distributions are possible to formulate we speak about Bayesian models

[7] Global Navigation Satellite System

3.3.1 Sequential Monte-Carlo – The particle filter

Mathematics inspired by nature

Sequential Monte-Carlo (SMC) methods are frequentist methods based on simulations which allow for computation of posterior distributions under non-linearity and non-Gaussianity. They follow an interesting "Darwinian process" of survival of the fittest. The particle filter is a closely related algorithm, known by this name as the simulations can be compared to particles, which will be "filtered" according to their goodness of fit.

Concept

Three main basic steps exist in the particle filter and can be seen in figure 3.4:

- sampling
- predicting
- measuring

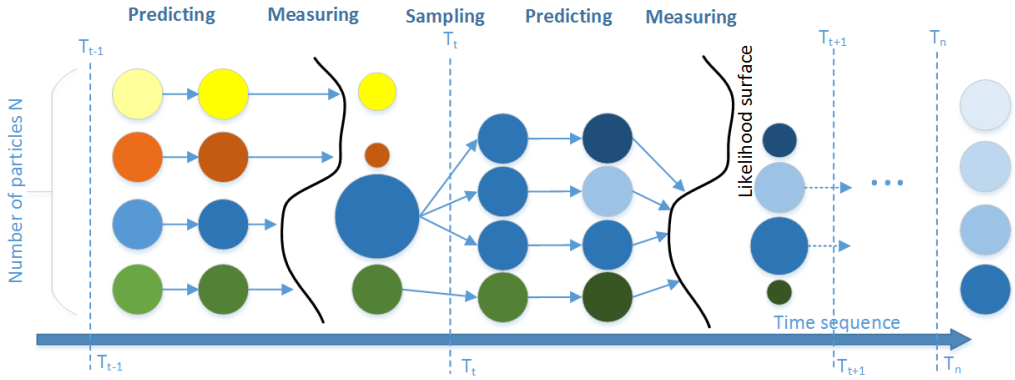


Figure 3.4: Scheme of the particle filter showing the 3 mains steps : sampling, predicting and measuring

The **sampling** generates N particles based on their weight. Particles with low weight will perish whereas particles with higher ones will be divided into several identical particles. As each particle is a simulation, it contains a set of parameters, meaning that N sets of parameters are generated.

These N particles (sets of parameters) then "evolve", they "predict" the next time step according to the model equations. It is the **predicting** step. As the model contains stochasticity, each particle will eventually evolve differently, even the ones that were identical after the resampling (it is the Monte-Carlo part, random processes are used, like in the gambling games played in Monte-Carlo). Each Monte-Carlo realization can be viewed as a particle's trajectory though the state space.

Afterwards the particles are compared to the observed data to get a likelihood (**measuring**). Their weight is then updated according to their goodness of fit[8].

[8] More precisely the weights are given according to the measurement model, which will be described later.

Finally the sampling occurs again, where the best particles are selected according to the posterior distribution. As mentioned, in order to avoid weight collapse and particle depletion, the resampling excludes particles with negligible weight and replace them in the proximity of particles with important weight. This also allows to focus the computations on regions of the parameter space with higher likelihood (and thus higher posterior probability).

These 3 main operations are performed sequentially for each time step until the end of the time sequence, thereby making Monte-Carlo simulations sequentially, hence the name of the method *Sequential Monte Carlo*.

A Darwinian selection

To summarize one can simply think about the analogy with evolutionary biology, as it is here that the Darwinian idea of survival of the fittest enters: some particles are born in the sampling part, then they evolve, "mutate" randomly (prediction). However only the ones that are adapted (the ones that have a higher likelihood) will survive and be able to give birth to offspring similar to their parents (resample), whereas the others will perish. This elegant (and ruthless) non-parametric methodology has numerous advantages such as:

- it can be used for a broader range of distributions (e.g. non-Gaussian)
- can cope with non-linearity
- is parallelizable
- focus computations adaptively on favorable regions of state-space

Pseudo-code

[9] This subsection presents a short pseudo-code (1) for the implementation of the particle filter. First let's define the unobserved states (or unobserved Markov process) as $\{x_t; t \in \mathbb{N}\}$, $x_t \in \chi$, which are modelled as a Markov process of transition equation $p(x_t|x_{t-1})$. The observations $\{y_t; t \in \mathbb{N}^*\}$, $y_t \in \mathcal{Y}$ "are assumed to be conditionally independent given the process $\{x_t; t \in \mathbb{N}\}$ and of marginal distribution $p(y_t|x_t)$"[53]. N is the number of independent and identically distributed particles $\{x_{0:t}^{1:N}\}$ according to $p(x_{0:t}|y_{1:t})$, and w their importance weight.

Particle Filter pseudo-code

Although the pseudo-code speaks for itself, some operations are worth more details. Line 10 is where the SIR model enters into action. It computes the next state from the previous one (from the state transition distribution, hence the Markov property). The particles obtained are thus the representation of the posterior. On line 11 the importance weights are computed thanks to the measurement model[10]. The last part, on line 14, changes the particle distribution by resampling the particles according to their importance weight. It is at this step that particles "not adapted to their environment" will perish, refocusing the particle set on regions of the state space with higher likelihood. Finally the time step is updated, setting the stage for the next sequence.

[9] This subsection is derived from the Introduction to Sequential Monte-Carlo Methods by Doucet et al. [53] and the lectures of Dr. Jizhong Xiao from the City College of New York on Advanced Mobile Robotics.

[10] Basically the measurement model relates the observed data with the hypothetical real data, it will be described in the modeling chapter.

Algorithm 1 Particle filter

1: **procedure** INITIALIZATION
2: $t = 0$
3: *for n=1 to N do*
4: $x_0^{[n]} \sim p(x_0)$
5: *endfor*
6: $t = 1$
7: **procedure** PF(χ_{t-1}, y_t)
8: $\bar{\chi}_t = \chi_t = \emptyset$
9: *for n=1 to N do*
10: sample $\tilde{x}_t^{[n]} \sim p(x_t|x_{t-1}^{[n]})$ ▷ Evolve the samples (process model)
11: $\tilde{w}_t^{[n]} = p(y_t|\tilde{x}_t^{[n]})$ ▷ Compute importance weight (meas. model)
12: $\bar{\chi}_t = \bar{\chi} + < \tilde{x}_t^{[n]}, \tilde{w}_t^{[n]} >$ ▷ Insert
13: *endfor*
14: $\{x_t^{[1:n]}, w_t^{[1:n]}\} \sim \{\tilde{x}_t^{[1:n]}, \tilde{w}_t^{[1:n]}\}$ ▷ Re-sampling according to the weights
15: *return* χ_t
16: set t = t + 1 ▷ Iterate PF (lines 7-16) until the end of the time sequence

For a proper and detailed mathematical formulation of the particle filter, the interested reader is advised to refer to the excellent Introduction to Sequential Monte-Carlo Methods by A. Doucet, N. De Freitas and N. Gordon [53].

3.3.2 Maximum likelihood via Iterated Filtering - MIF

Maximum likelihood via Iterated Filtering (MIF) is a particle filter based method developed by Ionides et al. [54], which allows a convergence to a maximum likelihood parameter estimate.

A particle filter based fitting method

When the parameters to estimate are time-varying random variables, their estimation becomes a reconstruction of unobserved random variables, allowing thus standard filtering methods, such as the one presented above, to proceed. However when those parameters are constant in time, things gets more complicated. The main motivation behind iterated filtering is therefore to create a particle-filter based method applicable to the inference of time-constant parameters by likelihood maximization.

Concept

Introducing a perturbation

The main idea in this method is to make the parameters time-varying. The model is thus replaced with a similar one where the parameters take a random walk in time, they will "suffer" from a perturbation after the resampling. The intensity of this random walk decreases while the number of SMC iterations increase (by a cooling factor), which will lead to the final estimate of the parameters as the new model will converge to the original one.

Figure 3.5 provides a conceptual view of this method. As previously mentioned, the main difference between MIF and the particle filter is the additional variability, which is included after the resampling step and depicted as corrugated arrows in the figure. When this "pseudo particle filter" reaches the end of its time sequence, the particles are moved back to the beginning the sequence, thereby performing another SMC but with better initial estimates. The particle filter is iterated (in this model it is iterated 50 times

for the initial guesses and 300 times for the refinement), hence the name of the method: *Iterated Filtering*.

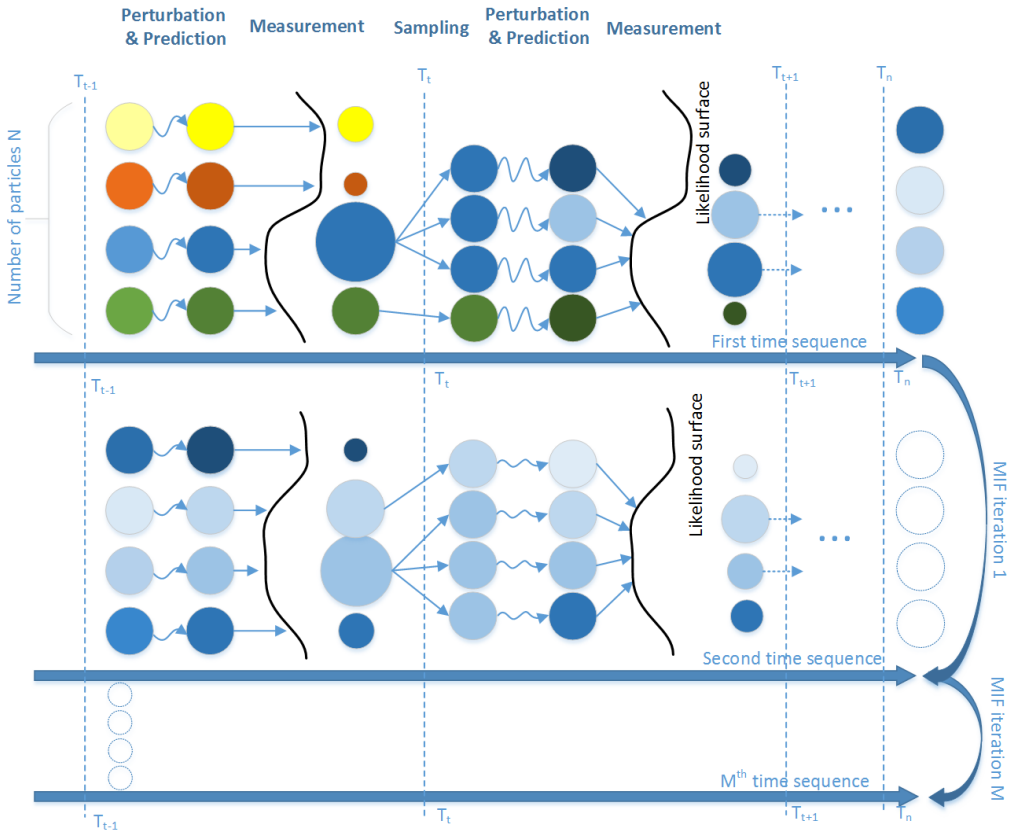


Figure 3.5: Scheme of the Iterated Filtering

According to the authors, introducing additional variability has three positive effects: "(i) it smooths the likelihood surface, which makes optimization easier, (ii) it combats particle depletion, the fundamental difficulty associated with the particle filter, and (iii) the additional variability can be exploited to estimate of the gradient of the (smoothed) likelihood surface with no more computation than is required to estimate of the value of the likelihood" [55]. MIF allows thus, efficiently, the inference of the parameters of nonlinear mechanistic dynamical systems with measurement error and covariates.

Pseudo-code

[11]In this subsection a simple pseudo-code (2) is presented in order to understand the basis of the iterated filtering. First let's define the time constant parameters θ as the time variant process θ_t taking a random walk in $\mathbb{R}^{d\theta}$. The densities $f(x_t|x_{t-1},\theta), f(y_t|x_t,\theta), f(x_0,\theta)$ are replaced by $f(x_t|x_{t-1},\theta_{t-1}), f(y_t|x_t,\theta_t), f(x_0,\theta_0)$. M is the number of SMC sequences, N the number of particles and T the end of the time sequence. The main algorithm depends only on the mean and the variance, which are defined as:

[11]This subsection is based on the *PNAS* paper by Ionides et al. [22]. The interested reader is advised to refer to this paper and especially the one published in *The Annals of Statistics* by Ionides et al. [54].

$$E[\theta_t|\theta_{t-1}] = \theta_{t-1} \quad and \quad Var(\theta_t|\theta_{t-1}) = \sigma^2\Sigma \tag{3.12}$$
$$E[\theta_0] = \theta \quad and \quad Var(\theta_0) = \sigma^2 c^2 \Sigma \tag{3.13}$$

Finally:

$$\hat{\theta}_t = \hat{\theta}_t(\theta, \sigma) = E[\theta_t|y_{1:t}] \tag{3.14}$$
$$V_t = V_t(\theta, \sigma) = Var(\theta_t|y_{1:t-1}). \tag{3.15}$$

MIF pseudo-code

Algorithm 2 Iterated Filtering

procedure INITIALIZATION
  $t = 0$
for n=1 to N do
  $\tilde{X}_{0,n}^{[F]} \sim p(x_0)$
endfor
  $t = 1$
procedure MIF
for m=1 to M do
  Set $\sigma_m = \sigma_0 \frac{(M-1)}{(M-2)+m}$ ▷ Hyperbolic cooling type
    for t=1 to T do
      Evaluate $\hat{\theta}_t^{(m)} = \hat{\theta}_t(\hat{\theta}^{(m)}, \sigma_m)$ and $V_{t,m} = V_t(\hat{\theta}^{(m)}, \sigma_m)$
      Apply PF$(\chi_{t-1}, y_t, \hat{\theta}_t^{(m)})$
    endfor
  Set $\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} + V_{1,m}\Sigma_{t=1}^{T} \frac{(\hat{\theta}_t^{(m)} - \hat{\theta}_{t-1}^{(m)})}{V_{t,m}}$
endfor
Return $\hat{\theta}^{(M+1)}$ as the maximum likelihood estimate of the parameter θ for the fixed parameter model

From the pseudo-code presented above one can see that the main difference with the particle filter is the introduction of a stochastic perturbation, which is applied to the parameters after each resampling.

This section closes the theoretical background of this thesis. The reader is now ready to enter the realm of cholera in the Bengal region of former British India.

4 | Cholera in Bengal - A historical dataset

4.1 Data

50 years of climatic, demographic, epidemiological records

A remarkably rich data set for the pattern of cholera epidemics, compiled by Dr Menno J. Bouma, exists for this region in the form of mortality records combining both urban and rural reports. The data was collected by the sanitary commissioners of the former British East Indian province of Bengal. It consist of monthly cholera death counts for 155 districts in 7 provinces from 1891 through 1941.

A decadal population census is available for the same period. The published results for 1891, 1901, 1911, 1921, 1931 and 1941 were used with linear interpolation after corrections for administrative changes. Moreover the temperature and rainfall data were recorded monthly by several meteorological stations per district, hence a monthly average was estimated for each location.

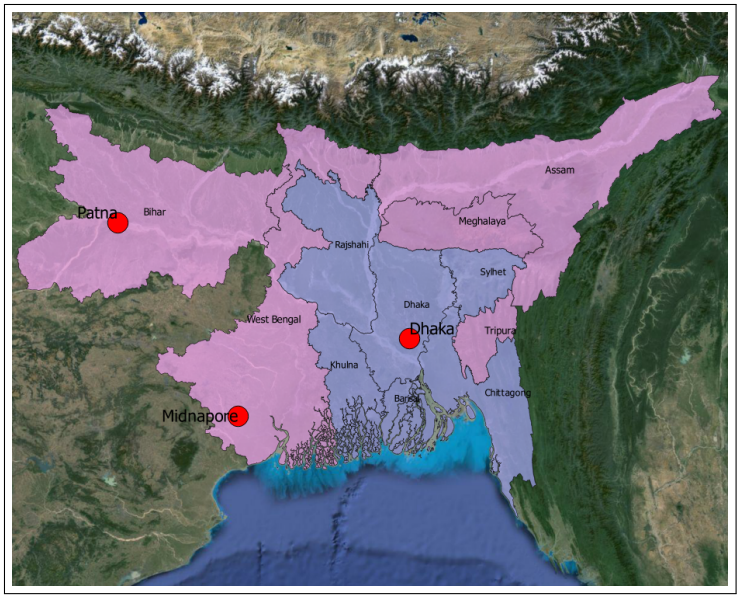


Figure 4.1: Limitation to the area under study to regions with at least 20 years of continuous epidemiological, demographic and climatic records. The red points correspond to districts detailed below.

4.2 Software

R and *QuantumGIS*

In order to process the data and identify the main patterns, two main softwares were used. The first one, *R*, is an open-source 4th level programming language and environment for statistical computing and graphics. It is based on the S programming language and is coded in C, C++ and Fortran. One of its strength is its modularity, thanks to the easy addition and creation of packages. Those numerous packages cover a very wide range of modern statistics [56] [1]. The second software, *QuantumGIS*, is an open-source Geographical Information System program and was used with its plugin OpenLayers to visualise the area under study and create the maps [57]. In addition to the aforementioned historical dataset, country-level data from DIVA-GIS has been used [37].

4.3 Disease and covariate patterns

Data patterns

In Bengal, 4 main regions with different patterns can be distinguished:

1. Midnapore is a district of the actual province of West Bengal. Its single annual cholera peak is representative of the dynamics found in the south-west coastal regions with pre-monsoon outbreaks during winter-spring. Interestingly it is one of the only regions where the median of the reported cholera cases is never null.

2. Dhaka, called Dacca during the colonial era, is the capital of Bangladesh and is located slightly more inland at the confluence of the Ganges, Brahmaputra and Meghna rivers. It is a low lying estuarine region which suffers from the famous double annual outbreak with infections pre- and post-monsoon. This seasonal pattern is the one observed in most regions of Bangladesh.

3. Patna is a north-western arid and dry area, it is one of the few places where a single annual peak, in phase with the monsoon, is observed. Being one of the most populated districts of the province of Bihar, it exhibits the dynamics found in that region.

4. Assam is a state located north-east. Cholera seasonality is more irregular in that area, which can almost be qualified as epidemic, most likely due to its lower population (see chapter 3.1). Although some results of the model will be shown, its dynamics will not be detailed in this chapter as they are less relevant.

Figure 4.2 below presents an overview of this diversity with the cholera patterns during the era of British India (1890-1940), whereas figure 4.3 shows the associated covariate patterns.

The median seasonal patterns of the first figure correspond to the Classical strain, which has been replaced over the years by El Tor. Nowadays, in Bangladesh, the dominant peak has shifted to the fall (September-October), and with the appearance of El Tor, the winter peak is observed almost 2 months before the one of the Classical strain [6]. Still this study remains valid for the present period as it has been suggested that "the habitats of the present form of cholera, El Tor, and the previous form, Classical, are the same. Conversely, El Tor risk areas during two time periods are not the same, which suggests that the present form of El Tor cholera may have changed environment and its habitat is now the same as Classical cholera before that biotype disappeared" [58].

[1] see section 5.3.1 detailing the POMP package used in this thesis

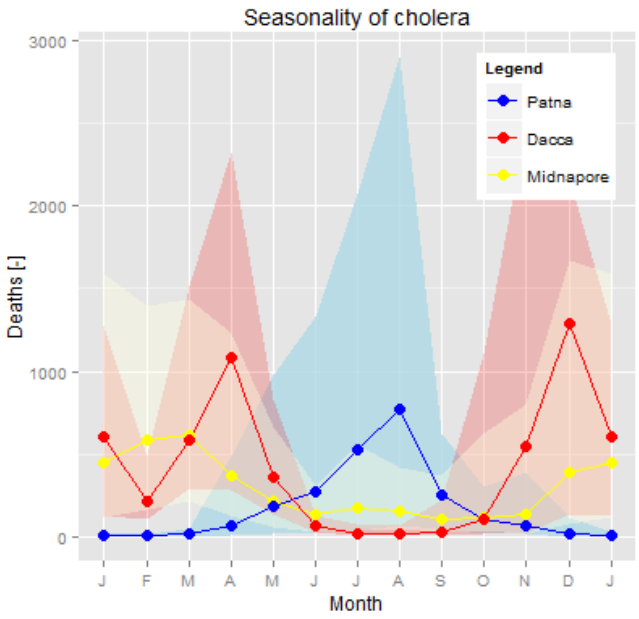
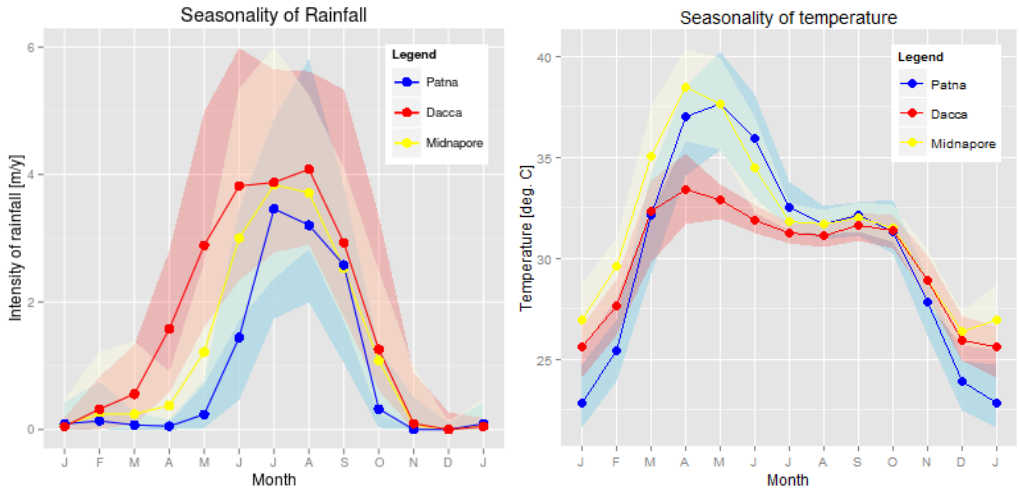


Figure 4.2: Three representative cholera seasonal patterns in Bengal (lines are medians and translucent strips 90% confidence intervals of the data).

Finally it has also been shown that the Classical biotype has again regained territory in recent years [28].

Covariates



(a) Rainfall **(b)** Temperature

Figure 4.3: Seasonality of covariates (lines are medians and translucent strips 90% confidence intervals of the data).

Interestingly, the seasonal patterns of the covariates are considerably similar (fig. 4.3) given the cholera cases reported. This might suggest differences in other environmental conditions among the areas, such as the hydrological regime. Moreover on can also notice that unimodal cycle distribution can result in a bimodal cholera prevalence, an interplay that has been widely discussed by the literature and explained only with the help of additional variables and interactions.

5 | Methodology and Modelling

The base model of this study has been established by my predecessor, Léonard Evéquoz, during its master thesis at the University of Michigan in 2013. He did a great job researching the literature, the existing models and processes. His work represents a significant first step towards modelling cholera dynamics in this endemic area.

5.1 Modelling framework

SIRBV model

To conduct the study a 7 compartments SIR-based model has been developed. Figure 5.1 represent a diagram of this compartmental model, where in addition to the SIR compartments, H stands for the human population entering the system (births), M the individuals dying from cholera, V the water body of the area, B the amount of pathogen in the area, λ the force of infection, T the local temperature and R the rainfall. It is worth mentioning that another output not shown in the diagram exists, the population dying from natural death. H, R, and T are inputs of the model and M is the observed state (cholera deaths reported), those 4 elements are the measurements.

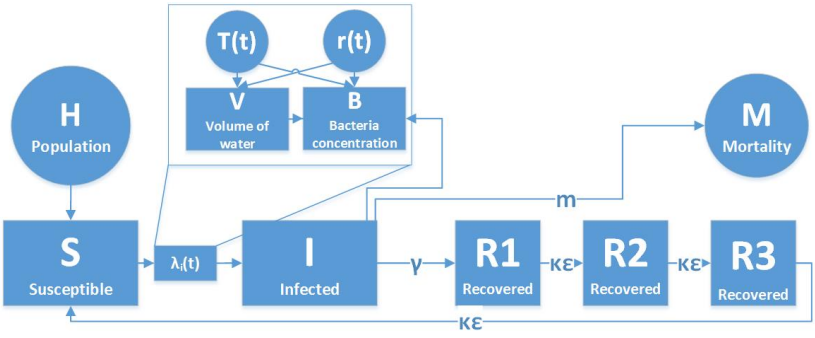


Figure 5.1: Diagram of the compartmental model

5.1.1 Model differential equations, states, and parameters

Process model

$S(t)$ is defined as the approximated real value of the susceptible pool at time t. $I(t)$, $R_1(t)$, ..., $R_k(t)$, $B(t)$, $V(t)$ are defined similarly. The diagram of figure 5.1 can be interpreted as the following set of coupled stochastic differential transition equations:

Process model

$$\frac{dS}{dt} = \kappa\epsilon R_3 + \delta H(t) + \frac{dH}{dt}(t) - (\lambda(t) + \delta)S \tag{5.1}$$

$$\frac{dI}{dt} = \lambda(t)S - (\gamma + m + \delta)I \tag{5.2}$$

$$\frac{dR_1}{dt} = \gamma I - (\kappa\epsilon + \delta)R_1 \tag{5.3}$$

$$\frac{dR_2}{dt} = \kappa\epsilon R_1 - (\kappa\epsilon + \delta)R_2 \tag{5.4}$$

$$\frac{dR_3}{dt} = \kappa\epsilon R_2 - (\kappa\epsilon + \delta)R_3 \tag{5.5}$$

$$\frac{dV}{dt} = r(t) - ET(T, V) - f(V) \cdot V \tag{5.6}$$

$$\frac{db}{dt} = -\mu_b(T)b + \theta(1 + \phi \cdot r(t))Ie^{-d(\bar{t}-t)} \cdot \xi(t) - f(V) \cdot b \tag{5.7}$$

With $f(V)$ as:

$$f(V) = \delta_i \frac{V(t)^{\alpha+1}}{V(t)^{\alpha+1} + \tilde{V}^{\alpha+1}} \tag{5.8}$$

and the force of infection $\lambda(t)$ modelled according to:

$$\lambda(t) = \beta \frac{\frac{b(t)}{V(t)}}{\frac{b(t)}{V(t)} + \bar{V}} \tag{5.9}$$

SIR states

S(t)

The **susceptible** population are the healthy individuals subject to a risk of contamination (eq. (5.1)). Its rate of change is composed of two inputs and outputs. The inputs are the births within the population (δH), with $\frac{1}{\delta}$ defined as the life expectancy (fixed to 50 years), and the people who are not immunized to cholera anymore, with κ the number of compartments and ϵ the inverse of the mean duration of immunity. The population size consists in an interpolation of the decadal population census. The outputs are the population getting infected ($\lambda(t)S$) and the one dying of natural causes (δS). The last term, $\frac{dH}{dt}(t)$, accounts for changes in the population (growth of decay) and can either be an input or an output.

I(t)

The **infected** compartment is the population affected by the bacterium (eq. (5.2)). The single input of this state variable is related to the force of infection and to the susceptible pool. Three outputs are considered, the people infected but dying from natural causes (δI), the ones dying from cholera (mI, with m being the fatality rate) and the ones recovering from the disease (γI, with $\frac{1}{\gamma}$ the mean duration of infection).

$R_k(t)$

The **recovered** compartments are used to model the people that, after an infection, acquired a non-permanent immunity. The outputs are the natural deaths (δR_k) and the waning of immunity (ϵR_k, where $1/\epsilon$ is the duration of immunity and κ the number of compartments) whereas the input is obviously related to the end of the infection (γI). As previously mentioned the duration of immunity remains unclear in the literature. While

some studies suggest an immunity lasting for several weeks [7], some others found a period in the order of years [50]. Several compartments are used in order to change the prior distribution of the individuals leaving the recovered state. As one compartment would result in an exponential distribution, the sum of exponential distributions will result in a gamma one, which allows a reduction of the variance as we have a prior idea of the recovery rate. The higher the number of compartments the smaller the variance, with an infinite number of compartments, for example, a constant output would be observed. The choice of three compartments (5.3)(5.4)(5.5) is a trade-off between efficient computation and reduced variance.

Non-SIR compartments - Environmental states

From chapter 3.1 it is seems quite clear now that the Bengal region dispose of an aquatic reservoir playing a significant role in the dynamic of the disease, whose nature is water-borne. Hence two state variables will be added to the SIR model, the water body of the area and the amount of pathogen present.

$V(t)$

The **water body** (eq. (5.6)), is a quantity in [m] ($[m^3/m^2]$). It is not the absolute quantity of water, it only informs about the "wetness" of the area. The hydrological regime of an area is usually linked to its meteorological conditions (rainfall, evapotranspiration), soil type (grounwater, runoff), topography and river network (discharge).

Rainfall

Although the importance of river discharge has been previously discussed this model only considers rainfall as water input in the area. Taking into account discharge would require a considerable amount of additional data as well as the hydrological network or topography, which is simply unrealistic for an area this size. Moreover a coupled approach would have to be considered in such a case, thereby increasing significantly the complexity of the model. Raw monthly rainfall data has thus been employed and interpolation used to satisfy the daily time step of the model.

Evapotranspiration

The evapotranspiration is considered as one of the two outputs of the water body. It is usually a function of temperature, soil type, vegetation and other meteorological conditions (radiative flux, humidity, wind). Evapotranspiration is computed according to the Blaney-Criddle formula [59]:

$$ET_p(T) = k_c ET_{p0}(T)[mm/d] \tag{5.10}$$

With:

$$ET_{p0}(T) = a(b\bar{T} + c)[mm/d] \tag{5.11}$$

Where $k_c = 0.8$, $a = \frac{N}{12 \cdot 365} \cdot 100$, $\bar{T}$ the mean monthly temperature, , and a and b coefficients detailed below. N is the mean daily percentage of annual daytime hours (function of the latitude (25° N in our case)).

Month	J	F	M	A	M	J	J	A	S	O	N	D
N (at 25° N)	10.7	11.3	12.0	12.7	13.3	13.7	13.5	13.0	12.3	11.6	10.9	10.6

Many formulas are available for computing such quantity: the Thornthwaite, Turc and Penman are other well known and widely used alternatives. However this formulation has one considerable advantage, it only requires temperature data (the others are function of global radiation, soil heat flux, water vapor pressure, wind speed, etc.). Nevertheless it

has a drawback, the evapotranspiration is significantly overestimated in non-arid/humid areas, which is the case of Bengal. To overcome this obstacle a "re-calibrated Blaney-Criddle" equation is considered. Weiss et al. [60] and Sperna Weiland et al. [61] propose new values for the b and c coefficients, more adapted to the study area (b = 0.35, c = 2.5).

Finally it is worth mentioning that this formula corresponds to a maximal quantity of water available valid only under the following conditions: (1) water is not a limiting factor, (2) the agronomic conditions are optimal (good fertility of the soil, sufficient nutrients, maximal vegetative development state, etc.). When the water volume in the area is low this formula is not valid anymore, hence the addition of the following statement:

$$ET(T,V) = \begin{cases} ET_p(T) \cdot \frac{V_i(t)}{T_{thresh}} & if \quad V_i(t) < T_{thresh} \\ ET_p(T) & else. \end{cases} \tag{5.12}$$

With T_{thresh} a threshold for normal evapotranspiration, in the model this parameter will no be fixed. This will allow the regions with different environmental conditions to have different evaporation behaviours. Some areas, such as the province of Bihar, are drier and more arid, the vegetation is scarcer and not under optimal growth conditions most of the year. One can thus expect a considerably lower evapotranspiration due to a low transpiration of the vegetation, thereby adding this flexibility can help catching those differences.

Drainage

The last component of the water body transition equation is the drainage. It corresponds to the water leaving the area, where δ_i is the drainage rate, and α and $\tilde{V}$ parameters allowing the drainage function to take different shapes (fig. 5.2).

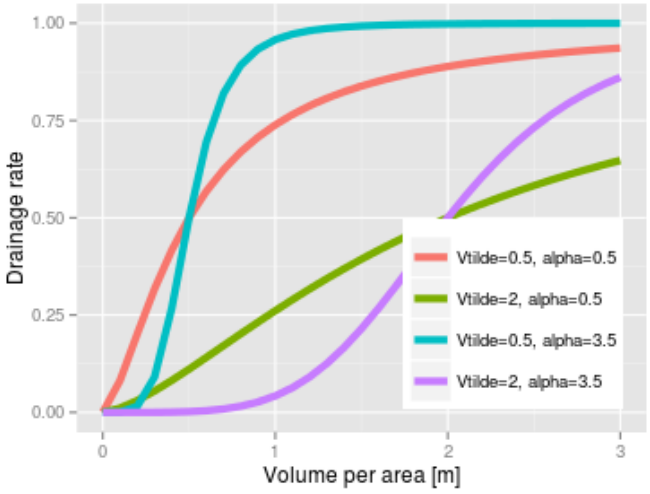


Figure 5.2: Behaviour of f(V) as a function of the volume and parameters ($\delta_i = 1$)

Figure 5.2 above illustrates the different behaviours of the drainage component. One can thus expect low elevation wet estuarine regions such as Dacca to follow more the blue curve, whereas northern drier area the red one. *f(V)* is multiplied by V to avoid having a maximal drainage.

$b(t)$

The last compartment is the amount of **pathogen** present in the area [-] (eq. (5.7)). Once again it does not represent the actual number of pathogen present in the environ-

ment, but provides an idea about the behaviour of this quantity[1]. This compartment considers a net death rate as output/input, an input due the pathogen shed in the environment by infected individuals, and a washout of the bacteria leaving the area through drainage. Among the several environmental variables influencing the growth/death rate of *Vibrio cholerae* (e.g. temperature, pH, salinity), temperature is the only data available in this historical dataset. Bertuzzo et al. [62] used the following formula to compute the bacterial death rate:

$$\mu_b(T) = \bar{\mu_b}(1 - \varepsilon \frac{T - \bar{T}}{T_{max} - \bar{T}}) \tag{5.13}$$

With the temperature in ° Celcius, $\bar{\mu_b}$ the average death rate of the bacterium (fitted parameter), ϵ the dependency to temperature (fitted parameter), $\bar{T}$ and T_{max} respectively the mean and maximal temperature of the area. It can be noted that this equation is simply a slope and could have been written in the form $y = m \cdot T + h$ with m and h parameters to fit. However the original formulation was preferred in order to have values comparable to the ones in the literature. Moreover in this model, the parameter ϵ is not bounded between 0 and 1 as proposed by the literature in order to allow the death rate to be a growth rate as well, thereby being referred to as a net rate.

Pathogen shed

The infected individuals shed pathogen in the environment which is represented by the term $\theta(1 + \phi \cdot r(t))Ie^{-d(\bar{t}-t)} \cdot \xi(t)$ of equation (5.7). Where θ is a parameter linked to the amount of pathogen per infected[2], and ϕ is a parameter allowing a positive effect of rainfall (bacteria could be washed out to healthy areas during rain events). The term $e^{-d(\bar{t}-t)}$ is an element allowing a decrease in the prevalence of the pathogen in the environment over time, $\bar{t}$ corresponds to the time at the half of the period. This additional term had to be added in order to fit regions having a strong decrease of reported deaths over time. Finally $\xi(t)$ is the process noise of the model, with $\xi(t) = \frac{dW}{dt}$ and $dW \sim \Gamma_{WhiteNoise}(0.015, dt)$. The process noise has been fixed to a value representing a good compromise between variability and a non-chaotic behaviour. A gamma distribution for the noise is chosen in order to avoid a negative white noise. Compared to other models where the process noise was introduced in the force of infection, better results were obtained with this new formulation. Indeed it provided more variability during important outbreaks in the reported cases, which resulted in better fits. The long-term trend change ($e^{-d(\bar{t}-t)}$) is as well usually introduced in the force of infection, as in chapter 3.2.1. Although it makes sense to reduce the contact rate over time (e.g. due to improved sanitation) a reduction of the amount of bacteria shed could also be observed (again thanks to improved sanitation for example). Moreover this formulation showed good results so it has been kept that way.

An unusual noise and trend

Washout

The last term accounts for the bacteria leaving the area within the water. It is not a concentration as the water body is only an indicator of the wetness of the area and the pathogen not the actual number of *Vibrio cholerae* in the environment.

[1] A normalization was applied to the actual bacteria quantity in order to reduce the number of parameters of the model (see footnote 2).

[2] More precisely due to a rearrangement of the original model in order to reduce the number of parameters, θ is the amount of pathogen per infected divided by the area of wetland, half saturation constant and an additional scaling parameter. For more details about transformation from the original model the reader is advised to refer to [63].

Force of infection

$\lambda(t)$

The force of infection links the susceptible pool with the infected one. β is the contact rate between the people and the infected water. Although speculations could be made as whether this parameter should be a function of temperature or volume of water and thus vary over time, it will be considered as a time-constant parameter here as few is known about such a relationship. Moreover it avoids increasing the number of parameters to fit. $\bar{V}$ is simply a scaling constant set to 1 as it doesn't affect the trajectory of I [3]. The force of infection is function of the water body and the pathogen, and can be assimilated to an idea of concentration.

Measurement model

Linking observations and simulations

The purpose of the measurement model is to relate the computed deaths with the observed ones y_n. It compares those quantities and return the likelihood of having those parameters given the observation, which corresponds to the probability density of obtaining the observation given the parameters ($\mathcal{L}(\theta|y_n) = P(y_n|\theta)$). Given a measurement data in a monthly time step, the number of cholera death for month n is thus $M_n = m\int_{(n-1)/12}^{n/12} I(t)dt$. The log-likelihood[4] is obtained through a negative binomial distribution as:

$$\log(\mathcal{L}) = log(NegBinom(y_n, \frac{1}{overdisp^2}, \rho M_n)) \tag{5.14}$$

Where *overdisp* is the dispersion parameter and ρ the reporting rate. The negative binomial distribution is a discrete dispersion that has the advantage of allowing some more overdispersion. Compared to the Poisson distribution, for example, the mean and variance can be different, which is especially interesting when the variance exceed the mean by a large amount. Practically it means that the model will not pay a prize too high in case it fails catching a peak in the data.

Parameters summary

15 parameters to fit

Below is a summary of the parameters fitted by the model:

- ϵ : Inverse of the duration of immunity $[y^{-1}]$
- γ : Inverse of the duration of infection $[y^{-1}]$
- T_{thresh} : Water height for a normal evapotranspiration $[m]$
- α : Non-linearity of the drainage $[-]$
- $\tilde{V}$: Shape parameter of the drainage $[m]$
- δ_i : Drainage rate $[y^{-1}]$
- $\bar{\mu}$: Normal bacteria death rate $[y^{-1}]$
- ε : Bacteria dependency to temperature [-]
- θ : Dose of pathogen per infected individuals $[y^{-1}]$

[3] Stating $b^* = b/\bar{V}$ and replacing it in equation (5.7), will simply results in writing $\theta^* = \theta \cdot \bar{V}$.

[4] As the logarithm is a monotonically increasing function, the maximum of the log-likelihood is reached at the same place than the likelihood, hence its use is preferred as products become additions and its derivative easier to compute when looking for its maximum.

- ϕ : Positive effect of rainfall $[y/m]$
- d : Decay of reported cases $[y^{-1}]$
- $overdisp$: Overdispersion parameter [-]
- ρ : Reporting rate [-]
- β : Contact rate $[y^{-1}]$
- m : Cholera fatality rate $[y^{-1}]$

5.2 Simulations and sensitivity analysis

The variability of cholera patterns in the Bengal region has already been discussed (cf. chapter 4). Therefore, before jumping into the more complex fitting procedure, a first approach has been considered in order to assess the flexibility of the model and its viability for the present study.

Chapter 4 showed that the covariates of the different regions are similar, moreover one can expect more differences in their hydrological parameters than in the epidemiological ones. Thereby 3 hydrological parameters of the model, δ_i (drainage rate), α (non-linerarity on the drainage) and $\tilde{V}$ (shape parameter of the drainage function), were selected as candidates for a sensitivity analysis.

A deterministic approach on MATLAB

To conduct the sensitivity analysis, a deterministic model has first been coded on the developping environment MATLAB, a 4th generation programming language developed by *The MathWorks* [64]. In order to reduce the computational time required, MathWorks's Parallel Computing Toolbox is used in order to run in parallel the simulations on multi-cores processors.

The simulations are realized by selecting a range of values for one parameter and fixing the others. The median annual rainfall and temperature for Dacca (red curves in fig. 4.3) are given as input and repeated every year until the model reaches a steady-state. A constant population is considered. A time-step of 1 day is chosen. For that reason a custom function based on the Dormand-Prince (DOPRI) method for solving the differential equations with a fixed and editable solving time step has been used. This solution has been chosen as MATLAB's ode45 function has an adaptive time-step. Finally an image displaying the infected population over a period of one year is crated, with each line of the image normalized by its maximum value.

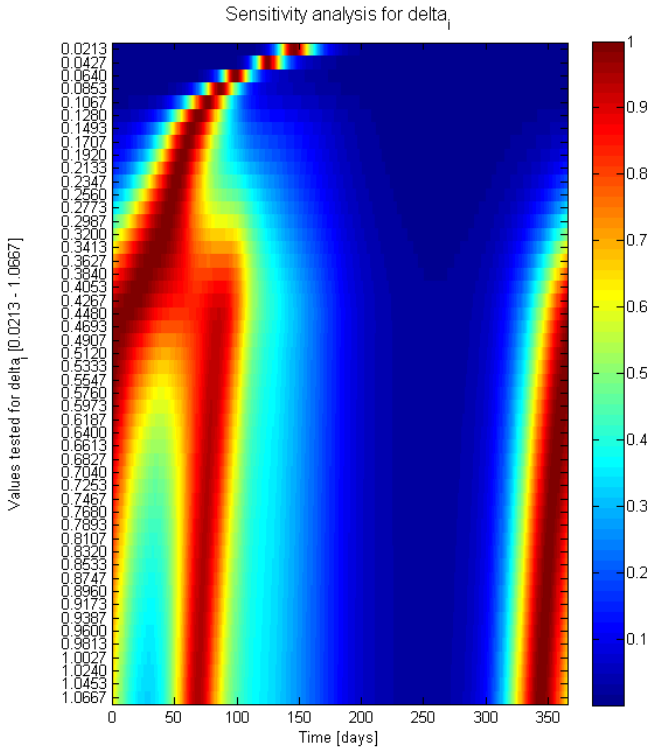


Figure 5.3: Infected population over a year for several values of the drainage rate δ_i (d^{-1})

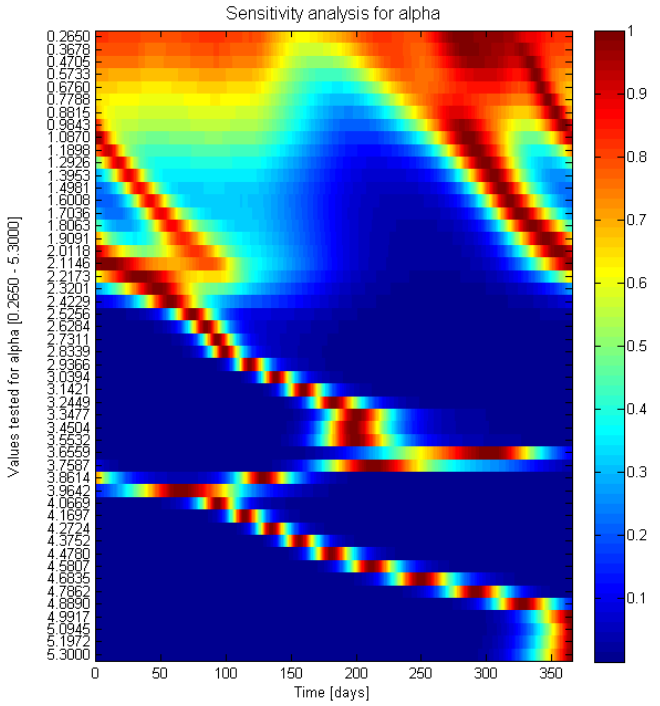
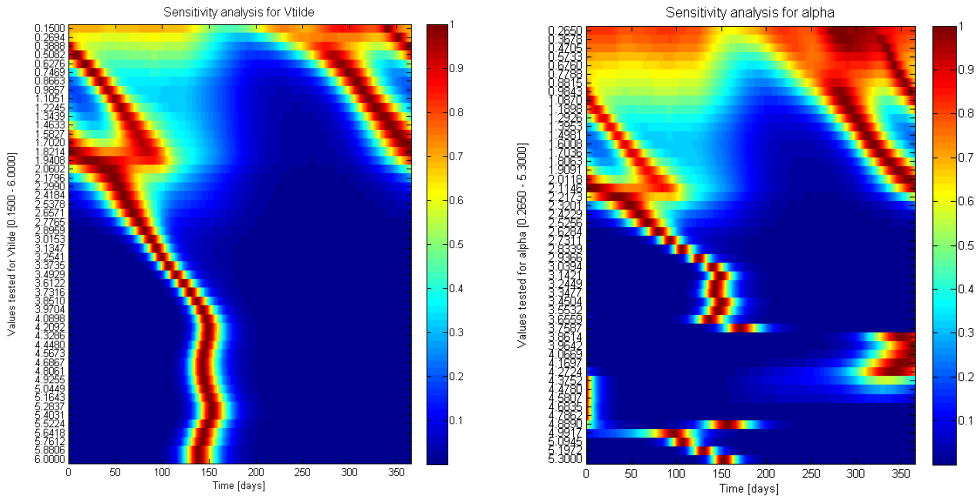


Figure 5.4: Infected population over a year for several values of α (-)



(a) Variation of $\tilde{V}$ (m) **(b)** Variation of $\tilde{V}$ and α

Figure 5.5: Infected population over a year for several values of $\tilde{V}$ and α

Dynamics shaped by hydrology

Figures 5.3, 5.4, and 5.5 reveal that by varying some hydrological parameters the model is able to produce a considerable range of seasonalities. Generally low values of α and $\tilde{V}$ seem to generate the double seasonal peak whereas higher ones tend to shift the pattern to an annual infection. The opposite pattern is observed with δ_i, the drainage rate. These results suggest that the model proposed has the required flexibility and is thereby an interesting candidate to model the wide range of cholera seasonalities in Bengal.

5.3 Computations

The implementation of the stochastic model has been entirely done with the use of the *Partially Observed Markov Process* package detailed in the next subsection.

FLUX HPC

Given the relatively computationally intensive fitting procedure, the model has been fitted on University of Michigan's High Performance Computing Cluster, FLUX, designed to support both compute and data intensive research [65]. This system is provided by ARC and operated by the College of Engineering's High Performance Computing Group within the Computer Aided Engineering Network (CAEN).

The computing procedure is the following:

1. Randomly generate 10,000 locations in the parameter space
2. Initial search: for each of those locations apply 50 MIF with 7500 particles
3. Refining: take the best 300 parameter sets from the initial search based on their likelihood and apply 300 MIF with 15,000 particles

At this point the plots can be made selecting some of the best sets based on their log-likelihood value. It is worth mentioning that computing this first part requires around 4 days on 200 computer cores.

To estimate the confidence intervals of the parameters, one of the best sets of the refining is selected, then the parameter to profile is fixed at several values, and a refining is performed with those new sets. The output is then divided into 20 regions of the parameter to profile. In each region the 3 best sets are selected and perturbed by $\pm 10\%$. All of those sets are then submitted again to the iterated filtering procedure which provides then the final profiles.

5.3.1 The POMP package

Implementing in POMP

Partially Observed Markov Processes (POMP) is a R package developed by King A. A. for fitting stochastic dynamical systems. Such models are composed of two parts, the true underlying process (unobserved process), which generates the data and the measurement process (observation process). The first step consists thus in writing both models in POMP. This is done with C snippets within the R code. In *Advanced Topics in POMP* [66], King A. A. shows that the C implementation results in up to a 120-fold speed-up compared to a standard and non-vectorized implementation in the R language. The use of a lower level language is thus considerably beneficial.

Parameters are sometimes subject to constraints (e.g. being positive). In order ensure that those constraint are satisfied it is sometimes advantageous to estimate them on a different scale than the one they appear in the model. Therefore most of the parameters of this model will be log-transformed within C snippets of the R code to satisfy those constraints (e.g. being positive in the case of a log-transformation).

Finally a pomp-object can be compiled and will be called by the by the mif function to proceed with the iterated filtering described in chapter 3.3.2.

For a more in-depth description of the functioning of the package, the reader is suggested to refer to the Introduction to POMP: Inference for partially-observed Markov processes by King A. A. [55].

6 | Results and discussion

6.1 Results

The model has been fitted for 6 districts within the study area. An overview of their location in Bengal is given by figure 6.1. The elevation map allows a situation in space of those locations and provides information about the environmental conditions one can expect in each of them:

- Dacca, a low elevation highly populated district surrounded by rivers and with important water bodies across its landscape.
- Patna, a populated arid north-western region.
- Midnapore, a populated near coastal district with an important decrease in population during the 1910's-1920's.
- 24-Parganas, a southern district located in the delta of the Ganges.
- Chittagong, a less populated eastern coastal area with a vast hilly terrain.
- Lakhimpur, a district part of the north-east corner of the state of Assam. It is a low population area subject to slightly epidemic cholera dynamics.

6 districts fitted

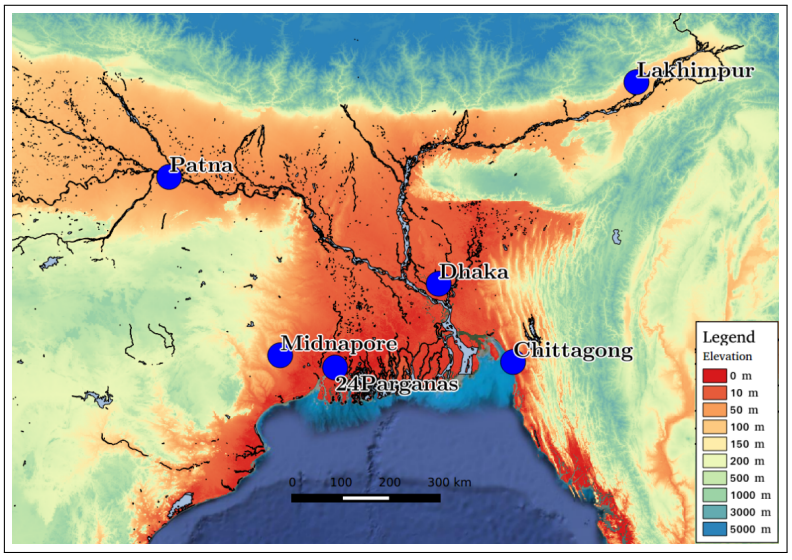


Figure 6.1: Elevation map of Bengal, circles indicate selected districts.

6.1.1 Seasonality

Seasonality in Dacca - The double peak

Figure 6.2 shows the median monthly cholera deaths for 40 years of data and simulation from 1900 to 1940. The curves exhibit the typical bi-modality of cholera observed in the capital of Dacca. The two pre- and post-monsoon peaks appear in spring and autumn, with a maxima in April and December, which corresponds to the Classical biotype pattern. The seasonality is captured by the model, with the peaks in phase with the data. The median of the simulations overlaps well the one of the data, except for a slight underestimation of the fall peak. This underestimation is not visible anymore when looking at the plots of the mean deaths in figure 6.4a instead of the median ones. The 90% confidence interval, however, catches this variability well, while it overestimates the winter drop and spring peak. The absence of deaths during summer is well captured by the model as well.

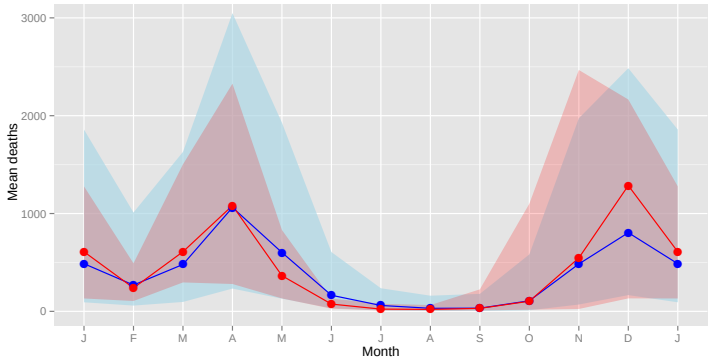


Figure 6.2: Cholera seasonality in Dacca. Median and 90% confidence intervals for cholera mortality, in red, and 250s simulations from the MLE model, in blue.

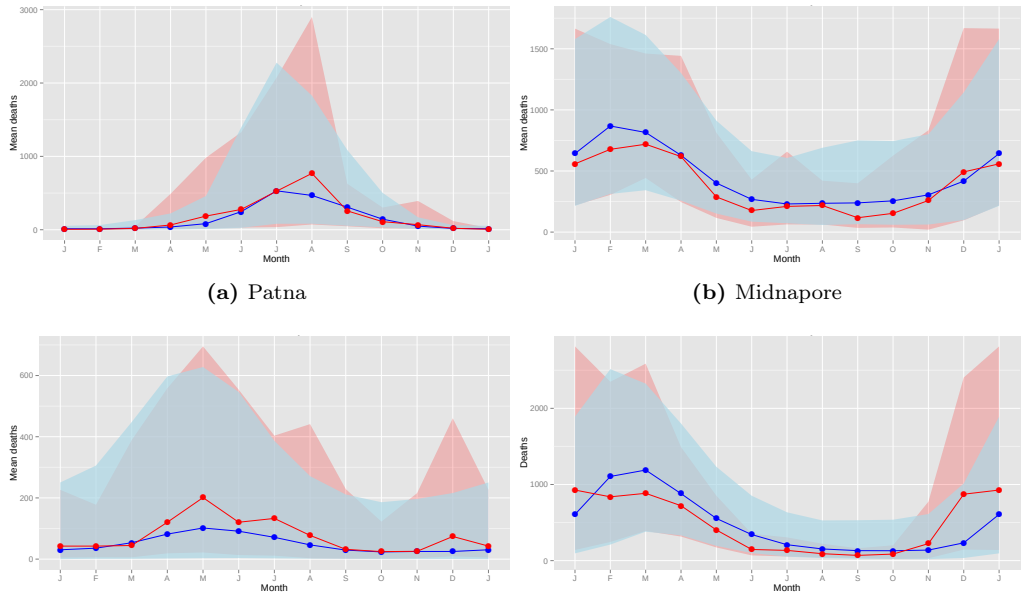
Single peaks

Figure 6.3 reveals also a good adequacy between the data and simulations. The single annual peak, observed during monsoon in the north-western drier region of Patna is captured (fig. 6.3a). The medians overlap well, apart from the month of August, where an underestimation is seen in the simulations. The confidence interval of the data has a more asymmetrical shape, with a sharp decrease after August, whereas the one of the model is more symmetrical. Finally while no cases are observed between January and March, some sporadic deaths are found in the simulations at that time.

The coastal district of Midnapore shows another interesting pattern: a single late winter-early spring peak (fig. 6.3b). Once again, it is caught by the model. Generally a slight overestimation is observed in the medians and in the confidence interval during summer. For the nearby district of 24-Parganas (fig. 6.3d) a similar pattern with a later but steeper fall-winter peak exists. In 24-Parganas however, the summer infection reaches lower values.

Chittagong (fig. 6.3c), a south-eastern district, presents a particular pattern: a main outbreak peaking in May followed by a small rebound once the monsoon rains appear. An irregular winter peak is revealed by the confidence interval but nearly non-existent in the median. This more random second peak is not captured by the model, whereas the first one, in spring, is underestimated by the median. Looking at the figure displaying the mean instead of the median (fig. 6.4b), one can see that the first peak is not underestimated anymore. However the rebound of July-August and the second peak of

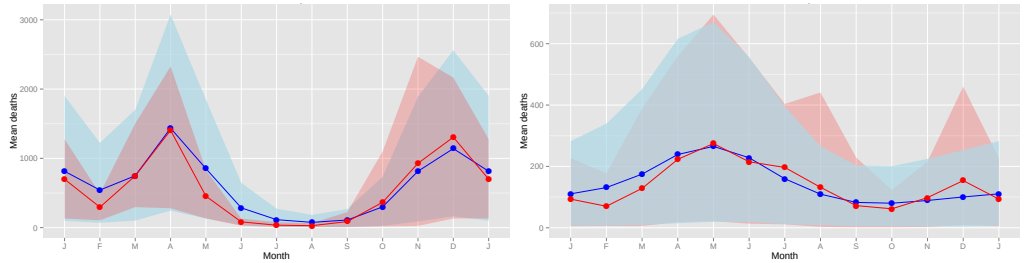
December remain uncaught.



(a) Patna **(b)** Midnapore

(c) Chittagong **(d)** 24-Parganas

Figure 6.3: Cholera seasonality for the districts of Patna, Midnapore, Chittagong and 24-Parganas. Median and 90% confidence intervals for cholera mortality, in red, and 250s simulations from the MLE model, in blue.



(a) Dacca (Mean) **(b)** Chittagong (Mean)

Figure 6.4: Cholera seasonality for the districts of Patna and Chittagong. Mean and 90% confidence intervals for cholera mortality, in red, and 250s simulations from the MLE model, in blue.

Northern areas, such as the district of Lakhimpur in the Indian state of Assam, present an interesting seasonality (fig. 6.5). The large confidence interval of the data with low median suggests a more irregular seasonality, which is close to epidemic dynamics although still showing signs of endemicity. These dynamics impact the catch of the seasonality by the model. The simulations are able to follow correctly the seasonal median of the data, however the confidence intervals after June do not overlap anymore. The fall bump of the confidence interval of the data is not reflected by the median, therefore suggesting more random infections and thus the inability of the model to reproduce this second softer outbreak. Due to its lower population, Lakimpur is the district showing the lowest number of cases, with the median peaking at 50 deaths per month.

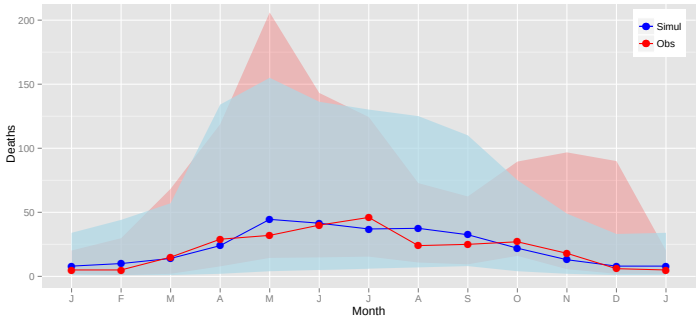
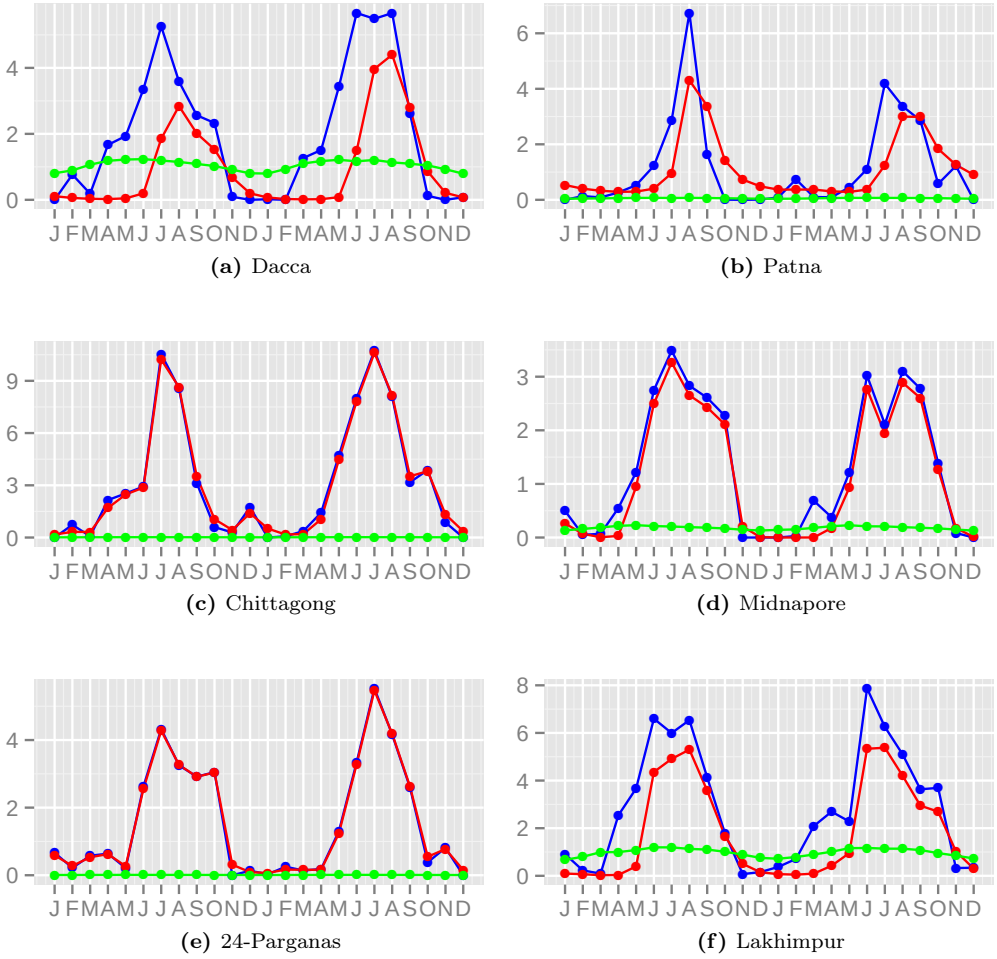


Figure 6.5: Cholera seasonality in Lakhimpur (Assam). Median and 90% confidence intervals for cholera mortality, in red, and 250s simulations from the MLE model, in blue.

Volume and bacteria components

Rainfall, ETp and drainage



(a) Dacca **(b)** Patna

(c) Chittagong **(d)** Midnapore

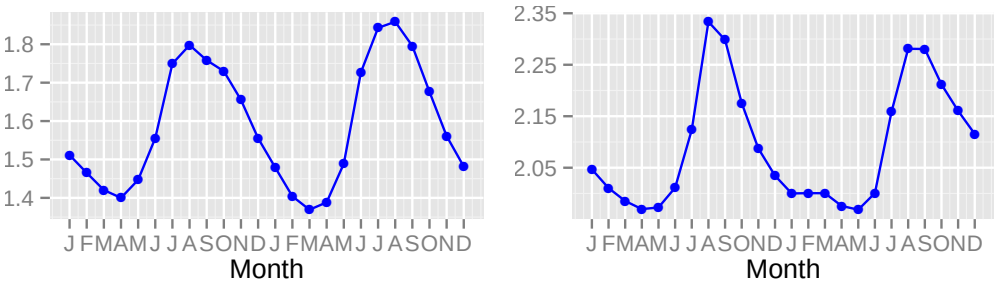
(e) 24-Parganas **(f)** Lakhimpur

Figure 6.6: Volume components over 2 representative years. Rainfall in blue, drainage ($f(V) \cdot V$) in red, and evapotranspiration in green [m].

Figure 6.6 shows the different components of the water body state (rainfall, evapotranspiration, and drainage) over a period of 2 representative years. Compared to the es-

tuarine region of Dacca, a near-null evapotranspiration in Patna is suggested by the model. More interestingly, the drainage has a slower response and the volume a longer persistence in Dacca. A 2 to 3 months delay and lower values are observed in this low elevation wetland-like area. In Patna, the drainage component persists longer after the rainfall events, hence suggesting a low water retention. A different pattern is observed in Chittagong and Midnapore, where the drainage curve follows the one of the rainfall. The model reveals once again a near-null evapotranspiration in Chittagong and low values in Midnapore. In terms of rainfall these districts are radically opposed as Chittagong presents the most intense precipitations whereas Midnapore the lowest ones. 24-Parganas, one of closest regions to the coast, located within the delta, follows the same behaviour than Chittagong. Finally Lakhimpur shows a pattern similar to Dacca with a high evapotranspiration, and a shorter delay in the drainage. However its rainfall events are significantly more important.

Water volume



(a) Dacca **(b)** Patna

Figure 6.7: Water body over 2 representative years.

The trajectories of the water volume in the districts of Dacca and Patna (fig. 6.7) are an illustration of the description provided in the previous paragraph. Dacca's low drainage rate results in a more eroded volume curve, whereas in Patna the water is discharged faster, leading to a sharper profile.

	Dacca	Patna	Midnapore	Chittagong	Lakhimpur	Parganas
Volume [m]	1.58	2.08	0.72	1.09	1.68	0.40
Bacteria [-]	0.0295	0.0172	0.0121	0.0062	0.0081	0.0062

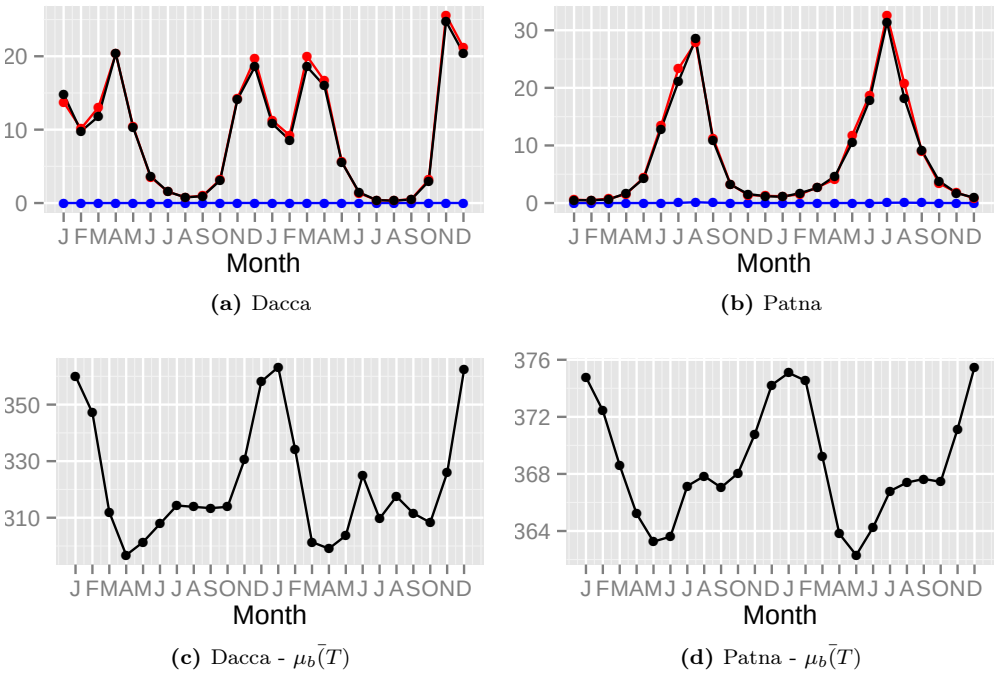
Table 6.1: Mean amount of water and bacteria for each district.

Table 6.1 shows the mean value of the non-SIR compartment for each region. Interestingly the highest values are found in the northern parts, with a maximum in Patna. The model seems to suggest lower values for the coastal districts (Midnapore and Chittagong). Regarding the bacteria, in order to reduce the number of parameters, the amount of bacteria has been normalized to a dimensionless number. The original formulation is the following:

$$b = \frac{b^*}{K A_c \bar{V}} \tag{6.1}$$

With b^* the "real" quantity of bacteria, K the half saturation constant, A_c the area of contact with contaminated water and $\bar{V}$ a reference depth. Therefore comparing the numbers given in table 6.1 cannot be done directly as the area of contact, different for each district, has to be taken into account. Unfortunately such a quantity is difficult to obtain.

Washout, death and input of pathogen



(a) Dacca **(b)** Patna

(c) Dacca - $\mu_b(\bar{T})$ **(d)** Patna - $\mu_b(\bar{T})$

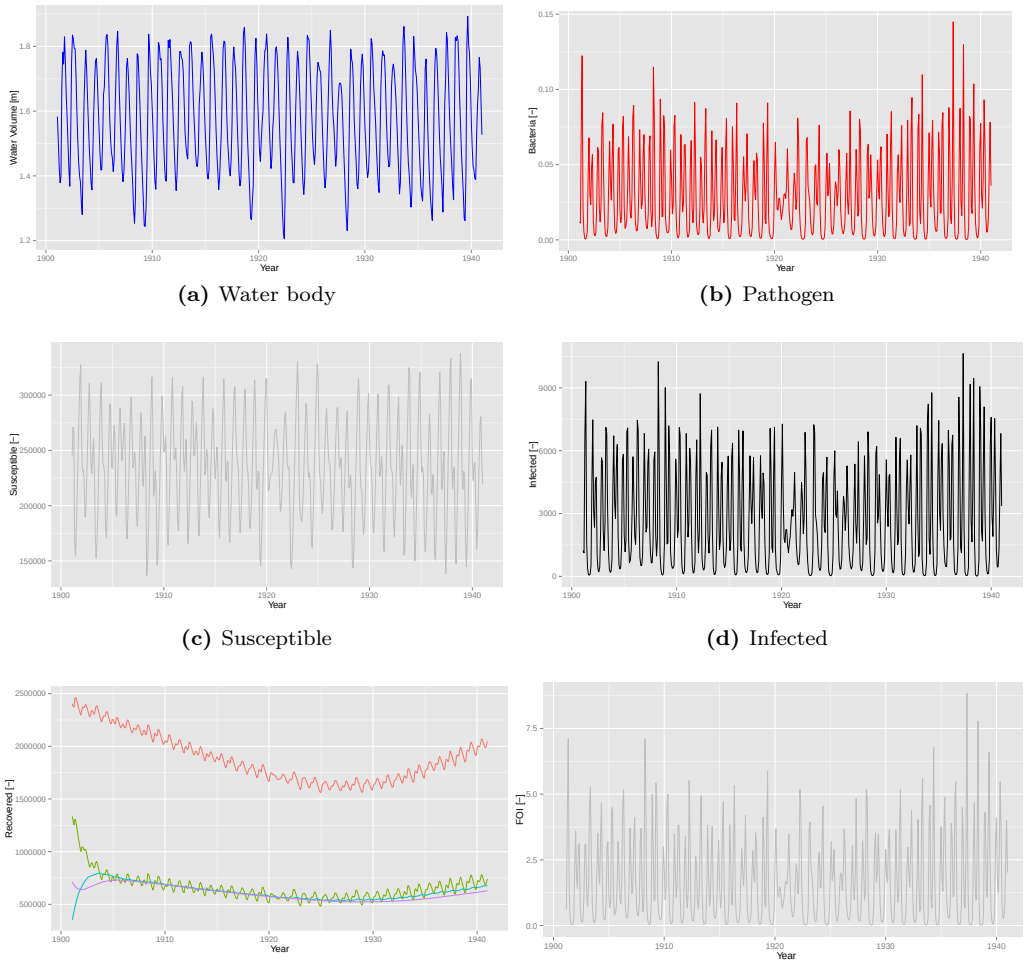
Figure 6.8: Pathogen components over 2 representative years. a) and b) washout in blue ($f(V) \cdot b$), input in red ($\theta(1+\phi r(t))I \cdot e^{d(\bar{t}-t}$), and death in black ($\mu_b(\bar{T}) \cdot b$) [-]. c) and d) death rate ($\mu_b(\bar{T})$) [y^{-1}].

The plots of the decomposition of the bacteria compartment[1] (fig. 6.8a and 6.8b) show the low importance in terms of values of the washout element compared to the input of pathogen and the death rate. Moreover one can see that the death rate is always positive although it is not positively constrained[2], therefore no growth seems to occur. The bimodal and unimodal patterns are clearly observable for Dacca and Patna in figures 6.8a and 6.8b as well. Figures 6.8c and 6.8d show the bacteria death rate for both districts. As expected maximum values are reached during winter, where temperatures are at their lowest. A lower death rate is suggested by the model in Dacca, whereas a lower dependency to temperature is found in Patna.

[1] Note that the noise is not included in plots 6.8a and 6.8b.
[2] Parameter b is not bounded between 0 and 1 (chapter 5.1.1).

Compartments time-series

Full time-series of the state variables



(a) Water body

(b) Pathogen

(c) Susceptible

(d) Infected

(e) Recovered (R1, R2, R3 in orange, grey and blue, R_{tot} in red)

(f) Force of infection

Figure 6.9: Full time-series of the state variables and the force of infection for Dacca (medians of 250 simulations).

Figure 6.13 reveals the full time-series of the SIRBV compartments and the force of infection simulated by the model. The water body, figure 6.9a, shows a steady sinusoidal pattern with oscillation between 1.4 and 1.8 m, and a mean around 1.6 meters. The pathogen, infected individuals, and the force of infection (fig. 6.9b, 6.9d and 6.9f) are subject to a similar pattern, which corresponds as well with the one of the simulated deaths (fig. 6.13a). The infected population is around 6 times higher than the fatal cases. The susceptible pool oscillates between 1,750,000 and 3,000,000 with a mean around 2,250,000 individuals. The recovered pool, however, seems subject to a depletion until the 1920's, where it starts to be replenished again. Due to their same parameters, the immunized people are evenly distributed among the three recovered compartments, with the only difference being an expected smoothed behaviour for the second and third ones.

6.1.2 Parameters estimation

Fitted parameters

The present table (table 6.2) provides a summary of the complete set of fitted parameters for each district after an initial search and refining. The log-likelihood, used to measure the goodness of fit, is also provided. Although it cannot be used to compare fits between different datasets, it can be used for a comparison with a similar models. The Akaike information criterion, another indicator of the relative quality of a statistical model, is given to compare models with different number of parameters[3]. The best previous model for the dataset of Dacca was developed by King et al. [7]. To provide an idea of the goodness of fit, a log-likelihood of **-3,050** was obtained. Although this result is considerably better it has to be kept in mind that it has been obtained with the semi-mechanistic model presented in section 3.2.1, in which the seasonality has been created artificially through the use of periodic cubic B-splines and therefore additional parameters.

	Dacca	Patna	Midnapore	Chittagong	Lakhimpur	Parganas
$1/\epsilon$ $[y]$	5.43	4.88	0.12	30.39	0.23	0.37
$1/\gamma$ $[d]$	3.01	1.20	1.22	1.86	1.03	1.01
T_{thresh} $[m]$	0.494	39.265	4.525	72.726	0.837	51.022
α $[-]$	18.85	13.96	154.72	52.64	21.88	101.67
$\tilde{V}$ $[m]$	1.87	3.41	0.75	1.10	2.21	0.41
δ_i $[y^{-1}]$	5.04	553.43	71.73	24.76	205.53	21.42
$\bar{\mu}$ $[y^{-1}]$	317.49	368.20	320.64	402.27	225.40	283.04
ε [-]	0.16	0.02	-0.06	0.04	0.04	-0.06
θ $[y^{-1}]$	0.0036	0.0877	0.0043	0.0380	0.0218	0.0047
ϕ $[y/m]$	0.0148	0.0298	0.0055	0.0028	0.0262	0.0079
d $[y^{-1}]$	0.0000	0.0072	-0.0013	0.0303	0.0211	0.0131
$overdisp$ [-]	0.82	0.52	0.32	0.48	0.53	0.13
ρ [-]	0.10	0.71	0.70	0.33	0.64	0.70
β $[y^{-1}]$	87.29	2.15	6.66	3.43	13.05	4.20
$mortality$ $[y^{-1}]$	23.70	42.12	7.29	51.43	7.58	23.56
loglik [-]	-3239.08	-2651.35	-3083.63	-2538.46	-991.41	-3115.04
AIC [-]	6508.16	5332.70	6197.26	5106.92	2012.82	6260.08

Table 6.2: Fitted model parameters of one of the best MLE models.

Hydrological parameters

5 hydrological parameters

The model considers 5 hydrological parameters, T_{thresh}, α, $\tilde{V}$, δ_i, and ϕ. T_{thresh} is a threshold for a normal evapotranspiration. In this model it allows an adaptation of the ETp to the conditions of each district. Eq. 5.12 shows that higher values of the parameter will tend to reduce evapotranspiration compared to low values. High evapotranspirations are observed in Dacca and Lakhimpur compared to lower ones in Patna and Chittagong. α and $\tilde{V}$ dictate the shape of the drainage function as shown in figure 5.2. Generally speaking, high values of both parameters will create a delayed drainage, which is what is observed in Dacca, Lakhimpur and Patna. The drainage rate, δ_i, corresponds to the maximum value this function can take. Patna is clearly ahead with a rate 100 times higher than the estuarine district of Dacca, hence suggesting more important water quantities leaving the area. Finally ϕ is the parameter related to the positive effect of rainfall, relatively higher in Patna where the single peak occurs during the monsoon, and in Lakhimpur.

[3]The AIC is based on the likelihood but penalizes a model for its number of parameters.

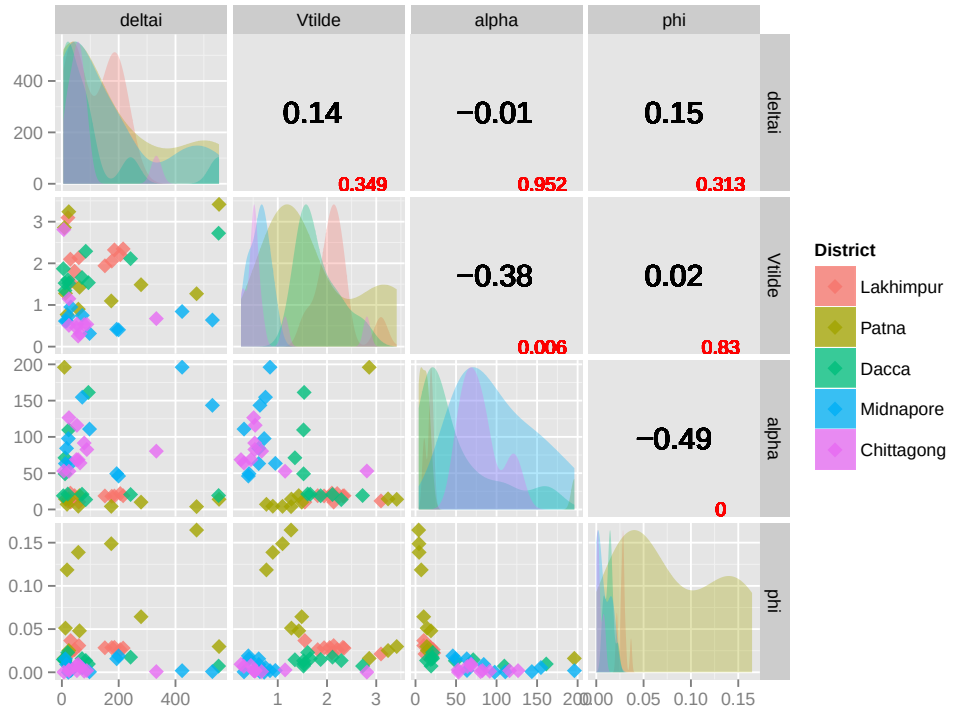


Figure 6.10: Hydrological parameters. Matrix plot for the 10 best sets of each location. Upper triangle: correlations in black and p-values in red, diagonal: scaled densities, lower triangle: scatterplots.

Figure 6.10 allows the same kind of comparison with higher number of parameter sets (the 10 best outputs of the refining of each district, which provide a more robust analysis). Moreover it informs about the correlations between the parameters. For the sake of clarity the district of 24-Parganas is not shown as its pattern is similar to the one found in Midnapore. From the densities, low values of $\tilde{V}$ with high ones of α are found in Midnapore and Chittagong, resulting thus in a high response on-off type of drainage function. High values of ϕ are found mainly in Patna. Finally significant negative correlations exist between α - $\tilde{V}$ (-0.38), and α - ϕ (-0.49).

Epidemiological parameters

5 epidemiological parameters

$1/\epsilon$, the duration of immunity, is subject to a considerable variation. Values from 44 days in Midnapore up to 30 years in Chittagong are returned by the model. The duration of infection, $1/\gamma$, is more constant across the districts (1.2 to 3 days). Mortality rates are up to 6 times higher in Patna and Chittagong than in Midnapore and Lakhimpur. β, the contact rate, is up to 40 times higher in Dacca than Patna. The mortality rate is particularly high in Patna and Chittagong, with minima in Midnapore and Lakhimpur. The lowest values of θ are observed mainly in Dacca and Midnapore. But once again, in order to reduce de number of parameters, θ doesn't correspond to the actual number of bacteria per people as it was proposed by the initial model:

$$\theta = \frac{p}{K A_c \bar{V}} \tag{6.2}$$

With p the amount of bacteria shed per people infected, K the half saturation constant, A_c the area of contact, and $\bar{V}$ a scaling parameters. θ is therefore in relation with the area of contact as well.

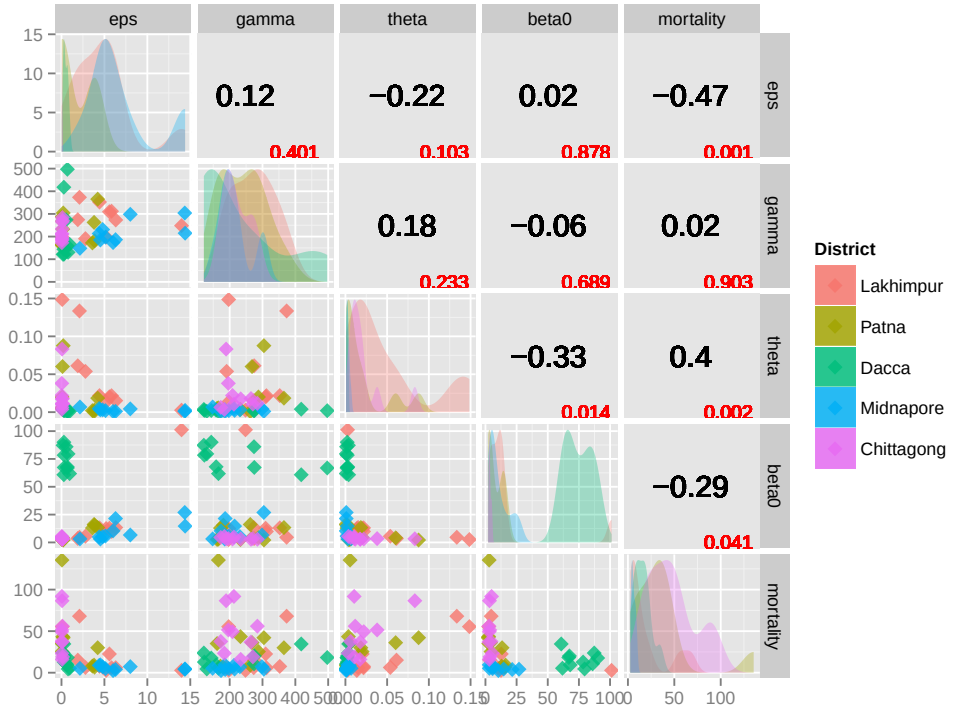


Figure 6.11: Epidemiological parameters. Matrix plot of the 10 best sets of each location. Upper triangle: correlations in black and p-values in red, diagonal: scaled densities, lower triangle: scatterplots.

The matrix plot (fig. 6.11) provides here similar results than the ones described in the previous paragraph. In addition to that a significant negative correlation can be noted between θ - β (-0.33), the duration of immunity with the mortality (-0.47), and the mortality with θ (0.40).

Other model parameters

ρ, d, *overdisp*, $\bar{\mu}_b$ and ε

The reporting rate, ρ, represents the percentage of reported death with respect to the ones that happened in reality. The highest values are obtained for the districts of Patna, Midnapore and Lakhimpur, where up to 70% of the deaths seem to be reported. A minimum is found for Dacca, the model suggests that the data represents only 10% of the fatal infections that occurred at that time. Chittagong suffers as well from a low reporting rate. Parameter d accounts for a long term change in the prevalence of the pathogen. This value is null for Dacca and near-null in Patna, whereas a decrease in time in Chittagong and Lakhimpur is observable in figure 6.13. Although its value is almost null, a slight increase is present in Midnapore. *overdisp* allows the addition of more variability in the measurement model than what would be expected. The bimodal pattern of Dacca requires the highest overdispersion values, being 1.64 times higher that the ones of the other regions. Midnapore presents the lowest one. Finally $\bar{\mu}_b$ and ε are

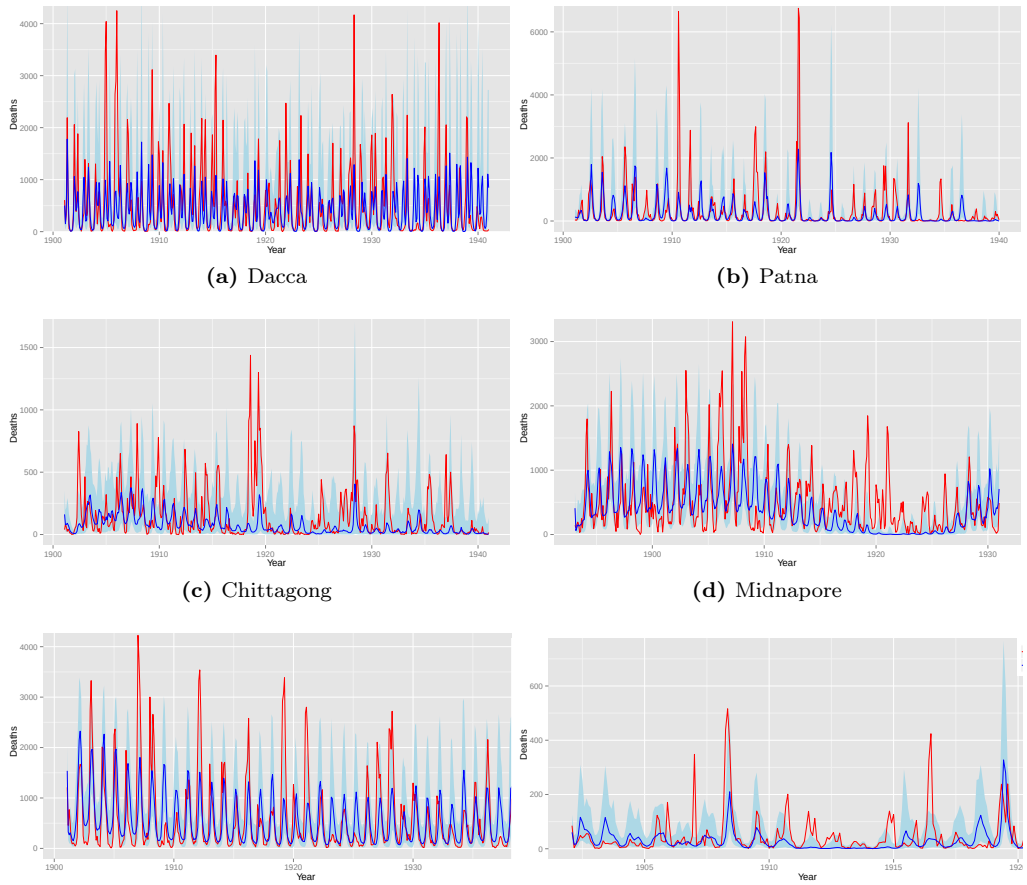
parameters acting on the death rate of bacteria. The average death rate under laboratory conditions for *Vibrio cholerae* of 70 y^{-1} is considerably lower that the ones suggested by the model for the Bengal region. The highest average death rate is found in Chittagong, the lowest in Lakhimpur. ε is a parameter commonly bounded between 0 and 1 in the literature and accounts for the dependency to temperature of the bacterium. Here it was completely free, which allows the death rate to become a growth rate if the value is higher than 1 or a negative dependency to temperature if it is negative. The district of Dacca presents the highest dependency to temperature, whereas interestingly in Midnapore and 24-Parganas a negative relationship is suggested.

Figure 6.12: Hydrological versus epidemiological parameters. 10 best sets of each location. Upper triangle: correlations in black and p-values in red, diagonal: scaled densities, lower triangle: scatterplots.

Figure 6.12 displays the eventual correlation between the hydrological and epidemiological parameters. Of particular interest are the positive and significant correlations between $\tilde{V} - \theta$, $\tilde{V} - \gamma$, and $\phi - mortality$ detailed in the discussion section.

6.1.3 Interannual variability

Interannual variability



(a) Dacca **(b)** Patna

(c) Chittagong **(d)** Midnapore

(e) 24-Parganas **(f)** Lakhimpur

Figure 6.13: Full time-series of the median of 250 simulations, blue, the 90% confidence interval, light blue, and the data in red.

Figure 6.13 shows the full time-series of the data and simulations. For the districts of Dacca and Patna, the median of the model captures partially but not fully the interannual variation. Dacca is more subject to frequent important outbreaks, few of them being captured by the median. Patna suffers less often from violent epidemics, however two of them, occurring in 1910 and 1921 are of particular intensity (up to more than 6500 monthly cholera deaths). Chittagong and Midnapore are subject to an important interannual variation, mostly followed by the model as well. The highly variable pattern of the district of Lakhimpur has already been mentioned. This results in a mediocre median fit, especially between 1911 and 1915. 24-Parganas is subject to a significant downward trend over the years, which is seen through the high value of its parameter d.

However, one can see that almost every outbreak is within the 90% confidence interval strip, suggesting that the model is capable of producing those behaviours, but that the stochasticity is determinant in their appearances. Figure 6.16, representing a unique simulation for every district, testifies of these stochastic dynamics, where a much more important variability and interannual variation is present. Finally, it is worth mentioning

that while the average deaths seem constant in time in Dacca, a slight downward trend is observed in Patna and Chittagong. Midnapore presents a curious variation due to an important decline of its total population during the 1910's-1920's.

Singular Spectrum Analysis and Fourier Analysis

To assess quantitatively the interannual variation of the data, Singular Spectrum Analysis (SSA), a statistical method allowing the decomposition of the time series into principal components, allowed to remove the seasonal component of the time series, hence extracting the interannual variation (fig. 6.14). Performing a Fourier analysis on the interannual component showed a periodicity of the anomalies in reported deaths of 4.2 and 7.8 years for Dacca (fig. 6.15a), 4 and 6 years for Patna (fig. 6.15b), and 7.8 years for Midnapore. Although in Patna this periodicity coincide well with the periodicity of the anomalies of the rainfall, generally no evident link between the periodicity of the anomalies in cholera deaths and rainfall or temperature could be found for the other regions.

Singular Spectrum Analysis

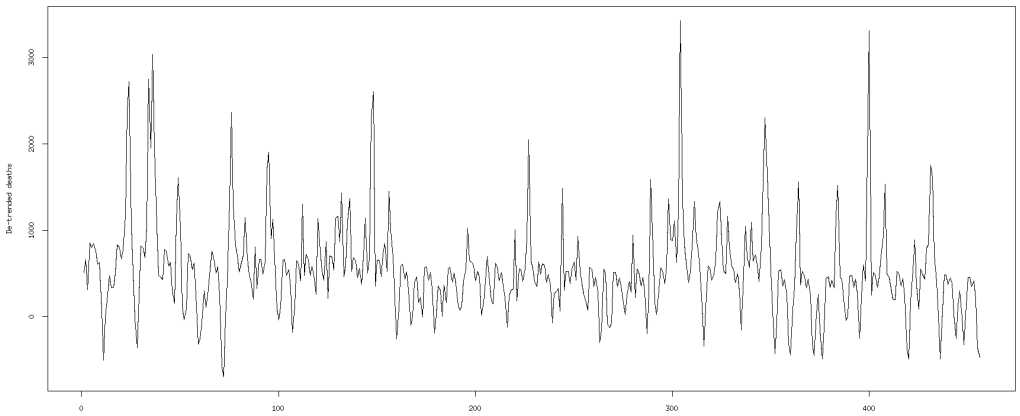
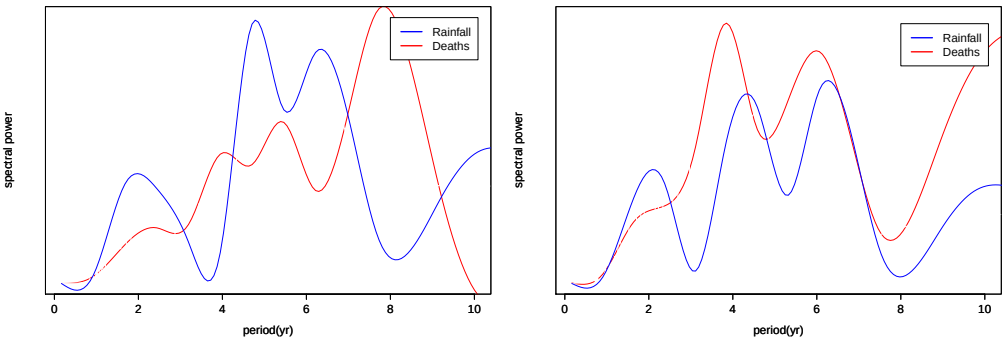


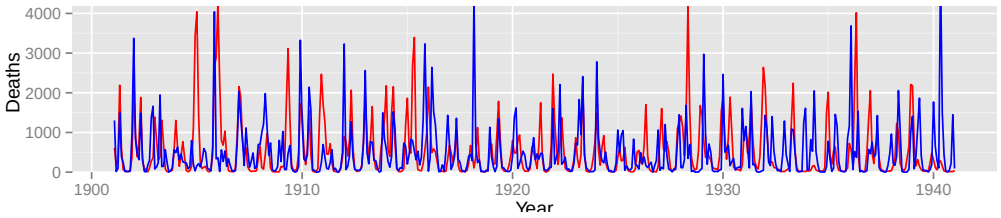
Figure 6.14: Singular Spectrum Analysis for Dacca with cancellation of the seasonal component.

Fourier analysis

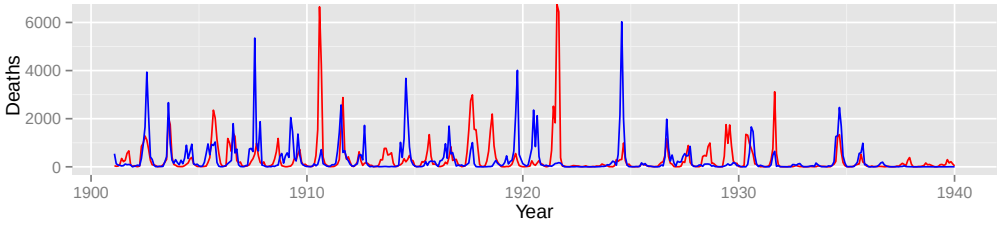


(a) Dacca **(b)** Patna

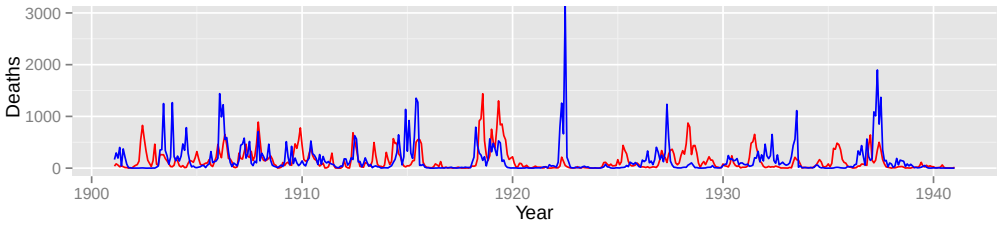
Figure 6.15: Fourier analysis on the anomalies of reported deaths and rainfall.



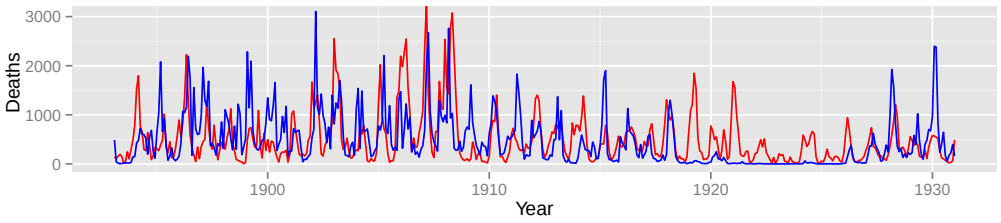
(a) Dacca



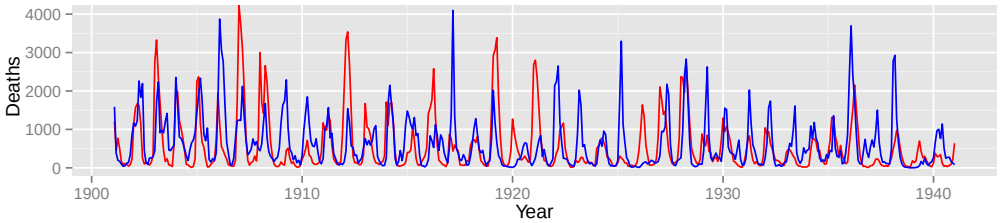
(b) Patna



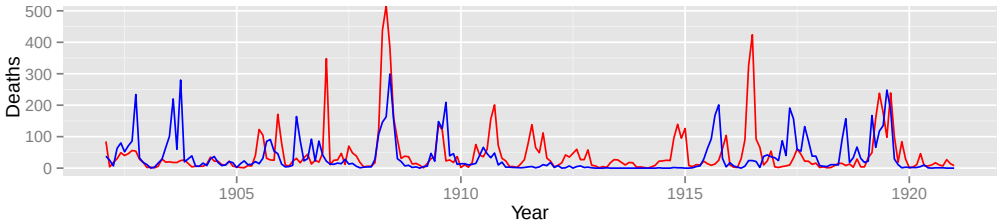
(c) Chittagong



(d) Midnapore



(e) 24-Parganas



(f) Lakhimpur

Figure 6.16: Full time-series of a unique simulation, blue, and the data in red.

6.2 Discussion

The results presented in the previous section revealed that a SIR model including a simple representation of the local hydrology is capable of capturing the wide range of seasonalities of the Bengal region. Using only rainfall and temperature as covariates with a single annual peak, uni- to bimodal distributions with annual infections pre-, post-, and during the monsoon were be obtained for various districts of the study area. All of the 6 regions assessed where at least partially, if not fully, fitted successfully. Moreover, in addition to the seasonality, the model showed a surprising ability to reproduce most of the interannual variation observed across the landscape.

6.2.1 Endemicity in Bengal - Mechanisms of seasonality

Double peak in Dacca

From the present study, it has been found that the first annual outbreak in Dacca is triggered by an increase in temperature and drier conditions in spring, which promote pathogen concentration and human contact rate due scarce water availability as suggested by Ruiz et al. [17]. Indeed, the bacteria death rate is, in this district, the most sensitive to temperature (highest ε returned by the model). The first peak is thus characterized by a strong primary transmission. These findings contrast with the explanations proposed by Akanda et al. [18], who hypothesize that the first peak is mainly modulated by coastal hydroclimatic conditions (SST, salinity, plankton abundance) and by the intrusion of salt water inland during periods of low river discharge (spring).

As suggested by Emch et al. [44], in Bangladesh, the important summer rains were shown to induce a dilution effect, presumably lowering the incidence of cholera. Akanda et al. emphasize that the peak streamflow observed in June creates important inundations spreading the pathogen across the landscape. Given the presence of water bodies in this estuarine region and the low drainage suggested by the model, this leads to large scale contamination, where the bacteria can thrive and develop without being washed out of the area (longer persistence). This sets the perfect stage for a new outbreak once the rainy season is over, the concentration of pathogen is higher, and the susceptible pool is replenished. Those results are in accordance with the more complex explanations provided by the literature stating that the important discharges of the monsoon decrease salinity levels and pH, but load the water sources with nutrients, hence favoring plankton blooms and bacteria development. Finally the decline in cholera infection observed in January and February is found to be temperature related, as suggested by Pascual et al. [6].

The drainage function in Dacca suggests that water remains in the area for a while after precipitation events. This phenomenon allows higher ETp rates (low T_{thresh} in the model) due to the creation of stagnant water bodies across the landscape (low drainage rate, δ_i), in accordance with the environmental conditions observed in this district. Being a wet district, Dacca experiences the dilution of the pathogen by rainfall (low ϕ), which lowers its concentration, leading to a decrease of the force of infection. Adding to these conditions, the much higher contact rate found by the model (β, more than 10 times higher on average than for the other regions), would lead to a higher sensitivity of the infection to the water reservoir and pathogen dynamics. Furthermore, even though the bacteria washout seems of low significance compared to its death rate, it still contributes to the equilibrium of the model, and its low values in Dacca most likely take part in the creation of a bacteria reservoir. Once the water level decreases, the concentration of bacteria increases again and the winter outbreak is triggered.

Interestingly the reporting rate (ρ) of 10% is the lowest of all values found for the different districts. This suggests a lower coverage or less efficient reporting system ,and therefore, that cholera fatality in this city was significantly higher than the reported values.

Single peak in Patna

Patna, with its single peak during the monsoon, shows a nearly opposite pattern. The hydrological regime of this area is of significant contrast with the capital of Bangladesh. Low evapotranspiration rates (high T_{thresh}) with faster and more intense responses of the drainage are found. Furthermore the persistence of high drainage rates even after the rain suggests a shorter retention time of the water in that area. This finding is consistent with what has been described in the literature. Patna is indeed a dry and arid district, in accordance with the low ETp suggested by the model, as during the dry season (winter and spring) no evaporation takes place even with the relatively high temperatures due to low water availability. Being an arid region, Patna would also experience a scarce vegetation cover and therefore, low transpiration during summer as well, when the water bodies are at full capacity. These low transpirations and low vegetation cover imply the more important drainage component (highest δ_i observed among all districts) described above.

The positive effect of rainfall in Patna (high ϕ) seems therefore related to its dry environment, as after a rain event a washout of the bacteria will occur, which will then fill up the reservoirs. This may suggest that the sanitary conditions and infrastructures in this region during the early 20th century are poor and sensitive to monsoon flooding.

Midnapore and Parganas

The districts of Midnapore and 24-Parganas, two coastal regions close to each other, are subject to an annual outbreak during the dry season (winter-spring). This kind of dynamics is once again shaped by the hydrological regime. The table of parameters shows that the highest values of α and the lowest of $\tilde{V}$ are proposed by the model for both regions, hence indicating the same type of hydrological dynamics. As could be expected, most of the parameters of these two regions are similar. The values of T_{thresh}, the drainage rate, and the mortality, exhibit the main differences between the two. The higher drainage in Midnapore is responsible for the earlier winter peak, probably due to the water volume reaching faster lower levels. The reporting rate, of 70%, is the highest of the whole Bengal. Although it might be delicate to assess the meaning of these values with certainty, an hypothesis could be that in these most populated regions occupied by the British East India Company due to their optimal location for trade benefits, a more rigorous and efficient reporting system was established. Interestingly the dependency of the bacteria death rate to temperature is positive only in those regions, which suggests either an exceeding of the optimal growth temperature, or the interplay with other phenomena. Chapter 4 showed that Midnapore exhibited some of the highest temperatures of Bengal, and chapter 2.1 mentioned that, according to the literature, 26 °C is the optimal growth temperature for the Classical strain (versus 37 °C for El Tor). Environmental signals can modulate ToxT-dependent virulence factor expression in *Vibrio cholerae*, and ToxT-dependent transcription in the bacterium can be significantly reduced or eliminated by an increase from 30 to 37 °C. Another explanation for the observation of this positive correlation is the fact that, during winter-spring, the discharge of the delta reaches its minimum value, therefore favouring the intrusion of salt water and plankton inland in coastal regions, which fosters *Vibrio cholerae* flourishing (see chapter 3.1). A non-environmental explanation may as well be related to those dynamics, as when facing dry seasons, people tend to store infected water and use it as their only source, which can create a continuous infection during the whole season, as seen in figure (fig. 6.3d) for 24-Parganas. Parameter ϕ suggest that no positive effect of rainfall is observed in any

coastal area.

Chittagong

The seasonality in Chittagong is among the most complex of the whole Bengal, yet the model is still capable to mimic some of its dynamics. Indeed, it is highly variable in frequency and explosive character, and figure 6.13c, for the interannual variation, testifies of those dynamics. For this reason, this irregular pattern is difficult to assess. A characteristic that differentiates Chittagong from other coastal areas is its unusual hilly terrain with the most intense precipitations among the fitted districts. Those factors justify the drainage curve observed and the near-zero ETp. Finally, an excessively high duration of immunity of 30 years is found, which is significantly more than the values found in the literature (few weeks to 3-10 years). This can be hypothesized that, although the district is fitted successfully by the model, its meaning and interpretation might not be entirely representative of reality.

Lakhimpur

Lakhimpur is the district in the northernmost part of the study region and the least populated one. Low population promotes stochastic behaviour because of facilitated and more frequent fadeouts. Thereby, as hypothesized by the literature, more epidemic dynamics can be expected, which is visible in figure 6.13f. The hydrological regime of Lakhimpur straddles the one of both Dacca and Patna, with important ETp but a fast response of the drainage. Precipitation in this region is intense and it typically spreads over a long period. Indeed, the monsoons in the eastern part of Bengal are more abundant than in the western part. A possible hypothesis is that this pattern results in a better irrigation over the year, thereby generating a better vegetation cover than in Patna and thus more vegetation transpiration. But again, assessing the seasonality of a district with those irregular dynamics is less relevant here given the goal and scope of this study, focused on more regular, endemic, areas.

6.2.2 Dynamics shaped by parameters

Hydrology

T_{thresh}

T_{thresh} is a threshold parameter for normal evapotranspiration. In the literature this parameters is usually fixed to 1 cm, which is considered to be a normal threshold for a maximal evapotranspiration. However, for reasons mentioned in chapter 5, it is not fixed in this model. Interestingly the model provides for most of the regions values higher than this boundary, which seems due to the fact that, in order to satisfy a steady-state (a null annual water balance), the ETp during winter has to be low in the regions with high drainage as no water is left. In turn, this results in a near-null ETp in summer. This behaviour seems to suggest a lack of flexibility of the re-calibrated Blaney-Criddle formula for those dynamics, and therefore jeopardize the use of this method to compute this quantity.

δ_i

The drainage rate, δ_i, has proven itself to be a parameter shaping the seasonality of cholera. The model showed that low values of drainage tend to create the double annual peak whereas high ones the single annual outbreak. Furthermore, as revealed in the sensitivity analysis (see chapter 5.2), and confirmed by the districts of Midnapore and 24-Parganas, lower values will have a tendency to delay the winter outbreak compared to higher ones due to the persistence of water (and hence dilution). Regions with lower drainage rates will therefore have longer water residence times, and although the dilution effect has a tendency to reduce the pathogenicity, low drainage will also tend to prevent bacteria from being washed out.

Epidemiology

Duration of immunity

Previous studies have found that infection-derived immunity to cholera ($1/\epsilon$) wanes on a timescale of 3 to 10 years. King et al. [7] recently found that the immunity even wanes from weeks to months. Those studies propose values in the range of the ones returned by our model, except for the district of Chittagong, which suggests an excessively high value of 30 years. However the particular dynamics of this region and its irregular seasonality already led to the conclusion that this model is most likely not fully adapted to that area. Still, observing durations of immunity varying from 6 week up to 5 years within the Bengal region seems considerably unrealistic, especially as no clear pattern is observable. Those results could be a consequence of the fitting procedure which considers every districts independently. To overcome this heterogeneity among the districts an optimal solution would be to fit several districts in parallel with the same epidemiological parameters but independent hydrological ones.

Duration of infection

The duration of infection, $1/\gamma$, is more uniform and values from 1 to 3 days are computed, which suggests a relatively short duration of infection. Dacca, the district subject to the double annual outbreak, with presumably the longest water residence time, shows the longest duration of infection. The correlations presented in the next paragraphs will provide information on whether or not this parameter is linked to the hydrological regime.

Fatality

The mortality rate is not homogeneous across the 6 fitted districts of Bengal. Midnapore and Lakhimpur are subject to low mortality rates compared to Patna and Chittagong. This might suggest a higher asymptomatic to symptomatic ratio. The implication is that most exposures do not result in severe cholera, but in mild or asymptomatic infections. However, because in this model all infected individuals are equally infectious, the vast majority of infectives could be assimilated to "silent shedders".

Correlations among parameters

Parameters correlated

The matrix plots of chapter 6 revealed several significant correlations (p-value < 0.05) within the set of parameters. One of the most interesting being $\tilde{V} - \gamma$ (0.30), which links hydrology with epidemiology. It suggests that the longer the delay of drainage (hence the longer the residence time of the water), the longer the infection. The hydrological regime is therefore in relation with cholera prevalence, as could be expected due to the waterborne nature of the pathogen. This corresponds to what is observed in reality, where *Vibrio cholerae* is frequently found in stagnant water bodies, which contribute to maintaining the infection. A strong correlation of 0.57 is found between $\tilde{V}$ and θ, suggesting that with a higher half saturation constant, the amount of bacteria shed per people is higher (for a same region, and thus area of contact A_c). The mortality rate is linked to several parameters. Indeed correlations 0.26, 0.40 and -0.47 are found respectively with ϕ, θ, and ϵ. Interestingly a shorter duration of immunity is related to higher fatality rates.

6.2.3 Cholera though the colonial era - The interannual variation

The stochasticity of abnormal outbreaks

The full time-series have shown that the model is capable of capturing some of the interannual variability of cholera only with rainfall and temperature. Although some particularly explosive outbreaks are not captured, most of them are within the 90% confidence interval of the 250 simulations. The contribution of stochasticity is thus significant. This suggests that either the model requires some adaptation or that the observed data is only a single realisation of a considerably stochastic natural process. The latter is what

is seen in figure 6.16, where non-averaged unique simulations show that the model is capable of reproducing the explosive outbreaks in terms of amplitude but at different times.

El Niño and anomalies

Nevertheless those abnormal outbreaks cannot be fully attributed to random events, as suggested by the Fourier analysis. Indeed the periodicity found in the anomalies (interannual component) of reported death imply an interplay with other climatic or demographic events. Interestingly enough, it roughly corresponds to the dominant frequency of El Niño (around 1/4 years), the most important phenomenon of interannual climate variability on a global scale. This is in accordance with the findings of articles published in *Science* [32] and *PNAS* [49], where the authors found that cholera dynamics are consistent with a remote forcing by ENSO (El Niño Southern Oscillation). Although this non-stationary link was mainly observed for the more recent datasets, ENSO certainly exerted an influence on the climate of the Indian Ocean during the colonial period as well (e.g. for example, after the warming of the Pacific, changes in cloud cover, evaporation, and increased heat flux can be observed a few months later in the Bay of Bengal, thus linking general climate to local variables impacting cholera).

The decline in deaths

Several districts such as Patna, Chittagong, Midnapore and 24-Parganas show a decline in time of the reported and simulated cases. Many hypothesis can be made to explain this long term trend, however not much can be done to assess them. One can for example explain it with a change of the reporting rate over time (changes in administration, demography, etc.). An improvement of sanitation in Bengal, reducing thereby cholera prevalence, is as well a valid one of the most likely explanations.

Population

Finally figure 6.13 reveals the stochastic and near epidemic behaviour of sparsely populated districts such as Lakhimpur and Chittagong. As a mentioned in chapter 3.1, the endemicity of an area is hypothesized to be linked to its human density. Indeed, Ruiz-Moreno et al. [17] showed that districts with a higher density have fewer cholera fadeouts compared to the ones with a lower population density. Hence stochastic patterns and thus epidemic dynamics are more prone to be observed in regions of the north-eastern parts of the Bengal.

6.3 Towards an improved modelling of cholera seasonality

This study is a second step closer to provide new insights into the seasonality of endemic cholera in the Bengal region. Encouraging results have been obtained with the first entirely mechanistic model based on rainfall and temperature. The journey, however, is not over and some future steps and directions are proposed in the following paragraphs in order to approach a denouement this thematic.

Global climate drivers

Although it was not the main goal of this project, improvements could be brought to the interannual variation by the inclusion of global climate drivers. The climatic interactions and influences at a regional level able to interact with the population dynamics of cholera in Bengal are numerous (e.g. snow melt in the Himalayas, SST, salinity level, etc.). However, as no records exist, few is known about the behaviour of those quantities in the early 20th century. Global climate drivers have the power to sidestep this difficulty as they encompass several of those variables into one. Being the most important phenomenon to cause global climate variability on interannual time scales, ENSO is a perfect driver accounting for such interactions. This has already been demonstrated by

Pascual et al., who found links between cholera dynamics and ENSO at the interannual scale. Moreover thanks to the availability of its records, ENSO would be an attractive candidate to include in the model. This improvement, however, would be made at the cost of having a less mechanistic model, as its interplay with cholera is not fully understood.

Confidence intervals

Due to time constraints for the present study, the profiling of the parameters could not be generated. It provides the confidence interval of a parameter, which allows then a more robust evaluation and analysis. It is therefore strongly recommended to proceed to the computation of the confidence intervals for at least some of the main parameters of each region.

A "parallel" fitting procedure

A problem that has been mentioned before is the important disparity observed within the values of a same parameter across the regions. Those contrasts are expected for the hydrological parameters, however few variation should be observed in the epidemiological ones. A suggestion to overcome this problem would be to fit some districts in parallel, where the epidemiological parameters of the districts will evolve together but the hydrological ones independently. This will result in attributing the variability of the seasonal patterns only to the hydrological regime instead of the epidemiology, which should be similar in the various districts of Bengal.

The coupled approach

Finally a more complex coupled approach could be considered. This coupled model will link the districts through a hydrological and a human mobility network. Chapter 2 showed that the Bengal region is highly interconnected by a complex river network. Using either simple river networks or more complex digital terrain elevation models, coupled models could allow taking into account snow melt, river discharge, and flooding, which are not represented in rainfall and temperature. Pathogen transport and spread might be modelled as well.

Human mobility is another driver playing a significant role in the interconnectivity of the districts and epidemiology of infectious diseases. Indeed while movements of infected individuals can propagate the pathogens across the landscape, displacement of healthy individuals can contribute to a replenishment of the susceptible pool, thereby extending outbreaks and epidemics. Unfortunately, the use of such models has some major drawbacks such as increased computational complexity and additional data requests. Therefore simpler fitting procedures will have to be considered.

7 | Conclusion

Cholera seasonality and the environment

For two hundred years, an explanation for cholera seasonality based on simple environmental drivers has remained elusive. Although the interplay between climate and diseases is being unravelled, and their interannual variation extensively studied, no clear and unified theory explaining the mechanisms governing endemicity has been proposed. Because the ecology of *Vibrio cholerae* and its transmission pathways (human-to-human and environmental-to-human) are still at the core of discussed unresolved mechanisms, this highly complex context led several scientists believe that simple environmental drivers such as rainfall and temperature, cannot explain it all. Through the present study, we provided the first evidences that a mechanistic rainfall-based model was capable of capturing the variety of cholera seasonal patterns, fully represented in the Bengal region.

Hydrology matters

Based on the results of our SIR-like model, with additional compartments for the water volume and the amount of pathogen, insights were gained on the conditions creating endemicity and its multitude of seasonal patterns. The hydrological regime proved itself to be a decisive driver determining the seasonal dynamics. Indeed, the present study found that rainfall tends to buffer the propagation of the disease in wet regions due to a dilution effect while enhancing cholera insurgence in dry regions. The more important drainage rate found in the dry district of Patna suggests higher discharges, possibly leading to sanitary infrastructures breakdown, hence boosting both primary and secondary transmission during the monsoon. Dacca, a wetter district with a double annual peak, is subject to delayed drainage responses, which promotes longer water residence times and longer duration of infection. Those completely opposite patterns indicate that overall water levels matter and appear to determine whether the effect of rainfall is positive or negative.

It is therefore undeniable that local environmental indicators, such as rainfall and temperature, are a key piece in understanding cholera seasonality. This finding contrasts with some of the outcomes revealed in the literature, where human-associated source of infections are often found more relevant to the dynamics of the disease than the cases related to the human-independent environmental reservoir. Furthermore, this modelling approach gave evidences that a persistence of the disease is provided by the environmental reservoir, which seems responsible for its endemicity.

Understanding cholera : the human benefits

Nowadays, with the advances in medical sciences, cholera is no longer the deadly burden it once was, yet it keeps causing despair in Bangladesh and many developing countries. Thereby, understanding the dynamics of the disease and being able to predict its behaviour could have tremendous human benefits worldwide. In this perspective, the mechanisms behind the seasonality of cholera have shown themselves to be in close relationship with climate and the environment, catching those mechanisms could thus allow better management and planning of public health policies in reaction to climate. Such

capacities are of paramount importance today, where behavioural changes of the population dynamics of infectious diseases in response to fast anthropogenic climate change will lead to new societal and scientific challenges in terms of disease prevention and mitigation strategies.

Finally, if the purpose of this study was to acquire a better understanding of the seasonality of cholera and gain insights on the causes of endemicity, it is ultimately oriented towards a contribution to cholera mitigation. In this regard, the use of such mechanistic models, complicated to implement as they require significant knowledge of the underlying natural processes, can be discussed as efficient alternatives inferring dynamics based on past behaviour and machine learning exist. One thing is certain however, its process-based nature, and the additional understandings it provides, brings us one more step towards the village of Gandhi's dream, in which *the villager will be all awareness, prepared to face the whole world, where there will be no plague, no cholera and no smallpox.*

Bibliography

[1] Barua, D. *History of Cholera*. Springer, 1992.

[2] Bryce, J., Boschi-Pinto, C., Shibuya, K., and Black, R. E. "WHO estimates of the causes of death in children". In: *The Lancet* 365.9465 (2005), pp. 1147–1152.

[3] *WHO | Diarrhoeal disease*. WHO. URL: http://www.who.int/mediacentre/factsheets/fs330/en (visited on 05/18/2014).

[4] *Cholera, World Health Organisation*. WHO. URL: http://www.who.int/mediacentre/factsheets/fs107/en (visited on 04/08/2014).

[5] *Epidémie de choléra à Haiti en 2010*. In: *Wikipédia*. Page Version ID: 99660263. May 2, 2014.

[6] Pascual, M., Bouma, M. J., and Dobson, A. P. "Cholera and climate: revisiting the quantitative evidence". In: *Microbes and Infection* 4.2 (2002), pp. 237–245.

[7] King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. "Inapparent infections and cholera dynamics". In: *Nature* 454.7206 (Aug. 14, 2008), pp. 877–880.

[8] Bertuzzo, E., Mari, L., Righetto, L., Gatto, M., Casagrandi, R., Rodriguez-Iturbe, I., and Rinaldo, A. "Hydroclimatology of dual-peak annual cholera incidence: Insights from a spatially explicit model: Hydroclimatology of cholera incidence". In: *Geophysical Research Letters* 39.5 (Mar. 2012), n/a–n/a.

[9] Emch, M., Feldacker, C., Islam, M. S., and Ali, M. "Seasonality of cholera from 1974 to 2005: a review of global patterns". In: *International Journal of Health Geographics* 7.1 (2008), p. 31.

[10] Akanda, A. S., Jutla, A. S., and Islam, S. "Dual peak cholera transmission in Bengal Delta: A hydroclimatological explanation". In: *Geophysical Research Letters* 36.19 (Oct. 10, 2009).

[11] Pascual, M, Chaves, L., Cash, B, Rodó, X, and Yunus, M. "Predicting endemic cholera: the role of climate variability and disease dynamics". In: *Climate Research* 36 (Apr. 30, 2008), pp. 131–140.

[12] Hashizume, M., Armstrong, B., Hajat, S., Wagatsuma, Y., Faruque, A. S., Hayashi, T., and Sack, D. A. "The Effect of Rainfall on the Incidence of Cholera in Bangladesh:" in: *Epidemiology* 19.1 (2008), pp. 103–110.

[13] Bertuzzo, E., Azaele, S, Martian, A, Gatto, M., Rodriguez-Iturbe, I., and Rinaldo, A. "On the space-time evolution of a cholera epidemic". In: *Water resources research* (2008).

[14] Gil, A. I., Louis, V. R., Rivera, I. N. G., Lipp, E., Huq, A., Lanata, C. F., Taylor, D. N., Russek-Cohen, E., Choopun, N., Sack, R. B., and Colwell, R. R. "Occurrence and distribution of Vibrio cholerae in the coastal environment of Peru". In: *Environmental Microbiology* 6.7 (July 2004), pp. 699–706.

[15] Emch, M., Yunus, M., Escamilla, V., Feldacker, C., and Ali, M. "Local population and regional environmental drivers of cholera in Bangladesh". In: *Environmental Health* 9.1 (2010), p. 2.

[16] Lipp, E. K., Huq, A., and Colwell, R. R. "Effects of Global Climate on Infectious Disease: the Cholera Model". In: *Clinical Microbiology Reviews* 15.4 (Oct. 1, 2002), pp. 757–770.

[17] Ruiz-Moreno, D., Pascual, M., Bouma, M., Dobson, A., and Cash, B. "Cholera Seasonality in Madras (1901–1940): Dual Role for Rainfall in Endemic and Epidemic Regions". In: *EcoHealth* 4.1 (Apr. 16, 2007), pp. 52–62.

[18] Akanda, A. S., Jutla, A. S., Alam, M., Magny, G. C. de, Siddique, A. K., Sack, R. B., Huq, A., Colwell, R. R., and Islam, S. "Hydroclimatic influences on seasonal and spatial cholera transmission cycles: Implications for public health intervention in the Bengal Delta: Hydroclimatic influences on seasonal cholera". In: *Water Resources Research* 47.3 (Mar. 2011), n/a–n/a.

[19] Altizer, S., Dobson, A., Hosseini, P., Hudson, P., Pascual, M., and Rohani, P. "Seasonality and the dynamics of infectious diseases: Seasonality and infectious diseases". In: *Ecology Letters* 9.4 (Mar. 31, 2006), pp. 467–484.

[20] Codeço, C. T. "Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir". In: *BMC infectious diseases* 1.1 (2001), p. 1.

[21] Siraj, A. S., Santos-Vega, M., Bouma, M. J., Yadeta, D., Carrascal, D. R., and Pascual, M. "Altitudinal Changes in Malaria Incidence in Highlands of Ethiopia and Colombia". In: *Science* 343.6175 (Mar. 7, 2014), pp. 1154–1158.

[22] Ionides, E. L., Bretó, C., and King, A. A. "Inference for nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18438–18443.

[23] Bhadra, A. "Time series analysis for nonlinear dynamical systems with applications to modeling of infecious diseases". PhD thesis. MI USA: University of Michigan, 2010.

[24] Nelson, E. J., Harris, J. B., Glenn Morris, J., Calderwood, S. B., and Camilli, A. "Cholera transmission: the host, pathogen and bacteriophage dynamic". In: *Nature Reviews Microbiology* 7.10 (Oct. 2009), pp. 693–702.

[25] DiRita, V. J., Neely, M., Taylor, R. K., and Bruss, P. M. "Differential expression of the ToxR regulon in classical and E1 Tor biotypes of Vibrio cholerae is due to biotype-specific control over toxT expression". In: *Proceedings of the National Academy of Sciences* 93.15 (1996), pp. 7991–7995.

[26] Schuhmacher, D. A. and Klose, K. E. "Environmental signals modulate ToxT-dependent virulence factor expression in Vibrio cholerae". In: *Journal of bacteriology* 181.5 (1999), pp. 1508–1514.

[27] Miller, C., Feachem, R. G., and Drasar, B. S. *Cholera epidemiology in developed and developing countries: new thoughts on transmission, seasonality and control.*

[28] Bouma, M. J. and Pascual, M. "Seasonal and interannual cycles of endemic cholera in Bengal 1891–1940 in relation to climate and geography". In: *The Ecology and Etiology of Newly Emerging Marine Diseases*. Springer, 2001, pp. 147–156.

[29] Acosta, C. J., Galindo, C. M., Kimario, J., Senkoro, K., Urassa, H., Casals, C., Corachán, M., Eseko, N., Tanner, M., and Mshinda, H. "Cholera outbreak in southern Tanzania: risk factors and patterns of transmission." In: *Emerging infectious diseases* 7.3 (2001), p. 583.

[30] Jutla, A., Whitcombe, E., Hasan, N., Haley, B., Akanda, A., Huq, A., Alam, M., Sack, R. B., and Colwell, R. "Environmental Factors Influencing Epidemic Cholera". In: *American Journal of Tropical Medicine and Hygiene* 89.3 (Sept. 4, 2013), pp. 597–607.

[31] Ali, M., Emch, M., Donnay, J.-P., Yunus, M., and Sack, R. B. "Identifying environmental risk factors for endemic cholera: a raster GIS approach". In: *Health & place* 8.3 (2002), pp. 201–210.

[32] Pascual, M. "Cholera Dynamics and El Nino-Southern Oscillation". In: *Science* 289.5485 (Sept. 8, 2000), pp. 1766–1769.

[33] Huq, A., West, P. A., Small, E. B., Huq, M. I., and Colwell, R. R. "Influence of water temperature, salinity, and pH on survival and growth of toxigenic Vibrio cholerae serovar 01 associated with live copepods in laboratory microcosms." In: *Applied and Environmental Microbiology* 48.2 (1984), pp. 420–424.

[34] Singleton, F. L., Attwell, R., Jangi, S., and Colwell, R. R. "Effects of temperature and salinity on Vibrio cholerae growth." In: *Applied and Environmental Microbiology* 44.5 (1982), pp. 1047–1058.

[35] Jutla, A. S., Akanda, A. S., Griffiths, J. K., Colwell, R., and Islam, S. "Warming Oceans, Phytoplankton, and River Discharge: Implications for Cholera Outbreaks". In: *American Journal of Tropical Medicine and Hygiene* 85.2 (Aug. 1, 2011), pp. 303–308.

[36] *Bengal - Wikipedia*. URL: http://fr.wikipedia.org/wiki/Bengal (visited on 06/02/2014).

[37] *Free Spatial Data | DIVA-GIS*. URL: http://www.diva-gis.org/Data (visited on 06/09/2014).

[38] Evequoz, L., Rinaldo, A., Pascual, M., King, A., and Bertuzzo, E. "Cholera dynamics in an endemic area : Capturing the intraseasonal climate forcing". In: *University of Michigan - Swiss Institute of Technology* (2013).

[39] Worldbank. "Water and Sanitation program. Long term sustainability of improved sanitation in rural bangladesh". In: (2011).

[40] *The World Fact Book*. URL: https://www.cia.gov/library/publications/the-world-factbook/geos/bg.html (visited on 06/02/2014).

[41] Ahmed, A. "Government of the People's Republic of Bangladesh. Bangladesh, climate change impacts and vulnerability, a synthesis". In: (2006).

[42] Colwell, R. R. and Spira, W. M. "The Ecology of Vibrio cholerae". In: *Cholera*. Ed. by Barua, D. and III, W. B. G. Current Topics in Infectious Disease. Springer US, 1992, pp. 107–127.

[43] Rahman, M. M., Hassan, M. Q., Islam, M. S., and Shamsad, S. Z. K. M. "Environmental impact assessment on water quality deterioration caused by the decreased Ganges outflow and saline water intrusion in south-western Bangladesh". In: *Environmental Geology* 40.1 (Dec. 1, 2000), pp. 31–40.

[44] Emch, M., Feldacker, C., Yunus, M., Streatfield, P. K., Thiem, V. D., Canh, D. G., and Ali, M. "Local environmental predictors of cholera in Bangladesh and Vietnam". In: *American Journal of Tropical Medicine and Hygiene* 78.5 (2008), pp. 823–832.

[45] Huq, A., Sack, R. B., Nizam, A., Longini, I. M., Nair, G. B., Ali, A., Morris, J. G., Khan, M. N. H., Siddique, A. K., Yunus, M., Albert, M. J., Sack, D. A., and Colwell, R. R. "Critical Factors Influencing the Occurrence of Vibrio cholerae in the Environment of Bangladesh". In: *Applied and Environmental Microbiology* 71.8 (2005). PMID: 16085859, pp. 4645–4654.

[46] Chun, J., Grim, C. J., Hasan, N. A., Lee, J. H., Choi, S. Y., Haley, B. J., Taviani, E., Jeon, Y.-S., Kim, D. W., Lee, J.-H., Brettin, T. S., Bruce, D. C., Challacombe, J. F., Detter, J. C., Han, C. S., Munk, A. C., Chertkov, O., Meincke, L., Saunders, E., Walters, R. A., Huq, A., Nair, G. B., and Colwell, R. R. "Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae". In: *Proceedings of the National Academy of Sciences* 106.36 (Aug. 9, 2009). PMID: 19720995, pp. 15442–15447.

[47] Colwell, R. R. "Global Climate and Infectious Disease: The Cholera Paradigm*". In: *Science* 274.5295 (Dec. 20, 1996). PMID: 8953025, pp. 2025–2031.

[48] Faruque, S. M., Naser, I. B., Islam, M. J., Faruque, A. S. G., Ghosh, A. N., Nair, G. B., Sack, D. A., and Mekalanos, J. J. "Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.5 (2005), pp. 1702–1707.

[49] Rodó, X., Pascual, M., Fuchs, G., and Faruque, A. S. G. "ENSO and cholera: A nonstationary link related to climate change?" In: *Proceedings of the national Academy of Sciences* 99.20 (2002), pp. 12901–12906.

[50] Righetto, L., Casagrandi, R., Bertuzzo, E., Mari, L., Gatto, M., Rodriguez-Iturbe, I., and Rinaldo, A. "The role of aquatic reservoir fluctuations in long-term cholera patterns". In: *Epidemics* 4.1 (Mar. 2012), pp. 33–42.

[51] Bellew, H. W. *The history of cholera in India from 1862-1881*. London: Tru?bner, 1885. 839 p.

[52] Roy, M., Bouma, M. J., Ionides, E. L., Dhiman, R. C., and Pascual, M. "The Potential Elimination of Plasmodium vivax Malaria by Relapse Treatment: Insights from a Transmission Model and Surveillance Data from NW India". In: *PLoS Neglected Tropical Diseases* 7.1 (2013). Ed. by Vinetz, J. M., e1979.

[53] Doucet, A., Freitas, N. d., and Gordon, N. "An Introduction to Sequential Monte Carlo Methods". In: *Sequential Monte Carlo Methods in Practice*. Ed. by Doucet, A., Freitas, N. d., and Gordon, N. Statistics for Engineering and Information Science. Springer New York, 2001, pp. 3–14.

[54] Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. "Iterated filtering". In: *The Annals of Statistics* 39.3 (June 2011), pp. 1776–1802.

[55] King, A. A., Ionides, E. L., Bretó, C., Ellner, S. P., KENDALL, B. E., FERRARI, M., LAVINE, M. L., and REUMAN, D. C. "Introduction to POMP: Inference for partially-observed Markov processes". In: (2012).

[56] *The R Project for Statistical Computing*. URL: http://www.r-project.org/ (visited on 06/09/2014).

[57] *The QuantumGIS project*. URL: http://www.qgis.org/en/site (visited on 01/04/2014).

[58] Ali, M., Emch, M., Donnay, J.-P., Yunus, M., and Sack, R. B. "The spatial epidemiology of cholera in an endemic area of Bangladesh". In: *Social science & medicine* 55.6 (2002), pp. 1015–1024.

[59] *Food and Agriculture Organization - Natural Resources Management and Environment Department*. URL: http://www.fao.org/docrep/s2022e/s2022e07.htm (visited on 06/02/2014).

[60] Wei\s s, M. and Menzel, L. "A global comparison of four potential evapotranspiration equations and their relevance to stream flow modelling in semi-arid environments." In: *Advances in Geosciences* 18 (2008).

[61] Sperna Weiland, F. C., Tisseuil, C., Durr, H. H., Vrac, M., and Beek, L. P. H. van. "Selecting the optimal method to calculate daily global reference potential evaporation from CFSR reanalysis data for application in a hydrological model study". In: *Hydrology and Earth System Sciences* 16.3 (Mar. 27, 2012), pp. 983–1000.

[62] Bertuzzo, E., Mari, L., Righetto, L., Gatto, M., Casagrandi, R., Rodriguez-Iturbe, I., and Rinaldo, A. "Hydroclimatology of dual-peak annual cholera incidence: Insights from a spatially explicit model". In: *Geophysical Research Letters* (2012).

[63] Baracchini, T. *Hydrologic controls on endemic cholera dynamics in Bengal region*. EPFL, 2014.

[64] *MathWorks - MATLAB and Simulink for Technical Computing - B*. URL: http://www.mathworks.com/ (visited on 06/11/2014).

[65] *Flux HPC Cluster | Advanced Research Computing at U-M (ARC)*. URL: http://arc.research.umich.edu/flux-and-other-hpc-resources/flux/ (visited on 06/11/2014).

[66] King, A. A. *Advanced topics in POMP*. University of Michigan, 2013.